Author:
Stadthagen, Hans

Title:
Age of acquisition and frequency effects in visual word recognition

AGE OF ACQUISITION AND FREQUENCY EFFECTS IN VISUAL WORD RECOGNITION

by

Hans Stadthagen

A thesis submitted in to the University of Bristol in accordance with the requirements of the degree of Doctor of Philosophy in the Faculty of Science, Department of Experimental Psychology.

April 2005

34,866 words

ALL MISSING PAGES ARE BLANK

IN

ORIGINAL

ABSTRACT

Age of Acquisition (A0A) and word frequency effects play a central role in the development of theories and models of visual word recognition. The main purpose of the present dissertation was assess whether these factors have independent roles in visual word recognition, and to provide a better characterization of each effect.

The first chapter provided an overview of the different hypotheses that attempt to explain AoA and word frequency effects. Chapter 2 presented the *Bristol Norms for Age of Acquisition, Imageability and Familiarity*, a large set of ratings that will facilitate the selection of experimental materials in the field. These ratings were used to replicate seminal findings concerning AoA and word frequency effects while addressing methodological concerns raised in the literature. It was found that both variables have a significant effect on word recognition when the other variable is matched across conditions.

Chapter 3 explored whether AoA and frequency have independent effects in visual word recognition, and particularly, whether AoA effects can be reduced to cumulative frequency effects or not. This was done using "expert vocabularies", words that were learned quite late in life but have very high frequency for a specific set of people. The results of these experiments are consistent with independent contributions of AoA and cumulative frequency.

Chapter 4 considered whether spoken word frequency has an effect on visual word recognition. Words with "unbalanced" frequencies (i.e., high spoken frequency but low written frequency or vice versa) were compared with two control sets in a lexical decision task. In both experiments, performance on the critical condition was similar to the control

set matched to its written frequency, and significantly different from the one matched to its spoken frequency. These results point towards a reduced influence of spoken frequency on visual lexical decision.

The implications of the findings from each part of the thesis were discussed in the light of current models of visual word recognition.

ACKNOWLEDGMENTS

Para Esmeralda, mi joya

I would like to express my deepest gratitude to my two supervisors, Dr. Jeff Bowers and Dr. Markus Damian, for their tireless support of my research and for their generosity with their time and energy. Through all my time in Bristol they have been my teachers and mentors, but most of all they have been my friends. I would also like to thank Dr. Colin Davis for many interesting discussions about issues raised within the thesis, and for his help in the elaboration of the Bristol Norms. I hope one day I will be able to repay the debt of time and knowledge that I have with him.

To my family, my eternal gratitude for all your support and love: you have been my examples and the source of my strength. My gratitude also goes to my friends in Costa Rica, Houston and Bristol, who have been my family when I'm away from home. I have a little bit of Fernandez-Somarriba, of Fernandez, of Castellanos and of Parraga in me. Thank you for accepting me into your families.

Finally, many thanks to all the participants that took part in the experiments presented in this thesis, particularly the faculty members at the different universities that took part despite their busy schedules and multiple obligations.

I would like to dedicate this thesis to my wife, Esmeralda, the love of my life, my jewel and my light. With love and wisdom you have helped me through tough times, you have been my shelter and my lighthouse, you have made sweet even the toughest of times. I thank God for letting me find you. AMDG.

ALL MISSING PAGES ARE BLANK

IN

ORIGINAL

AUTHOR'S DECLARATION

I declare that the work in this dissertation was carried out in accordance with the Regulations of the University of Bristol. The work is original except where indicated by special reference in the text and no part of the dissertation has been submitted for any other degree.

Any views expressed in the dissertation are those of the author and in no way represent those of the University of Bristol.

The dissertation has not been presented to any other University for examination either in the United Kingdom or overseas.

SIGNED: Hans Stadthagen

DATE: 5/10/05

TABLE OF CONTENTS

Section *Page*

LIST OF TABLES

Table Number *Page*

Chapter 1. Introduction

Chapter 1

INTRODUCTION

1.1. General Background

For most literate people, reading is a fast and automatic process that requires little conscious effort. However, this apparently simple operation is the result of complex processes that depend on a multitude of factors that span from low-level sensory analysis to mental functions of higher order, such as learning and attention. The question of how orthographic patterns (i.e., characters or words) are mapped into meaning or sound has attracted much attention from the cognitive research community over the last century. The focus on this issue stems not only from an intrinsic interest in how people read, but also because a better understanding of this process also has deep implications for how information is processed, stored and used by the brain in a more general sense. This knowledge is also important for applied fields, such as developing better techniques for teaching how to read, the study of dyslexia and other developmental language impairments, and the study and treatment of neurological patients with language loss or degradation.

A wide range of empirical studies, as well as theoretical and computational models have attempted to shed light into how printed words are recognized. Researchers have used multiple approaches to explore this issue, including studies with language-impaired patients (e.g., Coltheart, Patterson, & Marshall, 1987; Patterson, Marshall, & Coltheart, 1985), analyses of eye movements in reading (e.g., Rayner, 1998; Rayner & Sereno, 1994), brain imaging and electrophysiological techniques (e.g., Kutas & Van Petten, 1994), computational simulations (e.g., Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Seidenberg & McClelland, 1989) and, most relevant to the present dissertation, reaction-time studies of single word recognition (e.g., Taft, 1991).

One of the issues that has received much attention in the field is identifying those variables that influence the accuracy and speed of word recognition, as well as determining the relative importance and locus of the effect of such variables (for reviews, see Balota, 1994; Seidenberg, 1995). In this line of research, two factors have been the subject of much attention: how often a word is encountered (word frequency), and the age at which a word was learned (age of acquisition or "AoA").

Since very early on in the history of psycholinguistic research, word frequency has been considered one of the most important factors in determining performance in word reading tasks. Using a variety of paradigms, researchers have noticed that more frequent words are recognized more rapidly and more accurately than less frequent ones. Word frequency effects have been the focus of many studies dealing with the most varied aspects of language processing; indeed, word frequency plays a major role in virtually all current models of visual word recognition (for reviews, see Jacobs & Grainger, 1994; Monsell, 1991; Murray & Forster, 2004). Despite this large body of research into word frequency effects, there are still some issues that remain unresolved and some that remain understudied. The present dissertation will mainly focus on two such issues: the relationship between word frequency and age of acquisition (see below), and the degree to which spoken word frequency affects visual word recognition. A better understanding of these two issues will provide strong constraints for models of visual word recognition, as well as help researchers decide which variables should be controlled for in empirical studies.

In contrast to word frequency, the role of Age of Acquisition is less well established. In recent years, a growing number of studies have suggested that the age at which words enter the lexicon is also a major contributor to performance in word recognition tasks. According to these accounts, words that are learned earlier in life have an advantage in skilled performance over late acquired words. This factor has come to be called "Age of Acquisition" or "AoA". The status of AoA as an independent lexical variable is still controversial, mainly because of difficulties in disentangling the effects of AoA and frequency. It is particularly difficult to isolate the effects of these variables because they are highly intercorrelated; words that are learned at an early age tend to be more frequent through life, and conversely, less frequent words tend to be learned later. Research in the area has been further hindered by the fact that there are relatively few large norming studies, which are needed in order to control for relevant variables in factorial and regression designs; as can be seen in section index of norms presented in Chatper 2 of this thesis, there are only two AoA norming studies in English with more than 1,000 items each (Bird, Franklin, & Howard, 2001; Gilhooly & Logie, 1980a). As things stand at the moment, the issue remains unresolved and is the subject of a heated debate in the literature with regards to the relative importance of these two factors as well as the mechanisms by which they operate. .

A better characterization of AoA and word frequency effects is of great importance in the elaboration of theories and models of visual word recognition. The main purpose of the present dissertation is to advance our knowledge of these two effects (AoA and word frequency), firstly by providing evidence that both factors are relevant and have independent roles in visual word recognition, and secondly, by providing a better characterization of each effect. In the case of AoA, the discussion will be centered around the issue of whether AoA effects can be attributed to cumulative frequency (a measure of the total number of exposures to a word). In the case of word frequency, this thesis will consider an issue that, surprisingly, has remained largely unexplored in the literature: the degree to which spoken word frequency affects visual word recognition. Throughout the development of these main topics, other important issues related to AoA and frequency will be considered, such as the relationship between these two factors and other lexical variables, and the validity of word familiarity as a good representation of word frequency.

Section 1.2 of this chapter will consider some basic issues related to word frequency effects, and offer a brief review of how different models of visual word recognition have accounted them. It will also present some methodological issues about how word frequency estimates are obtained. Section 1.3 is dedicated to Age of Acquisition effects and will follow a similar structure: it will consider the different explanations that have been offered to account for AoA effects, and discuss different methods that have been used to estimate the age at which words are learned.

1.2. Word Frequency Effects in Visual Word Recognition.

As mentioned before, word frequency has long been considered the preponderant factor in determining the speed and accuracy of lexical access. Since very early in the word-recognition literature, there are references to an advantage in performance for words that are found more often in a language over less frequent words (i.e., *time* is recognized faster than *tame* and *face* faster than *foot*), highlighting the important role played by word frequency in lexical access. Reports of frequency effects can be found in early studies on reading of word-lists (Pierce & Karlin, 1957) and single-word recognition thresholds (Havens & Foote, 1963; Howes & Solomon, 1951; McGinnies, Comer, & Lacey, 1952). More recent studies have confirmed these results and extended them to other tasks such as single-word reading (Berry, 1971; Forster & Chambers, 1973; Frederiksen & Kroll, 1976; Grainger, 1990; Monsell, Doyle, & Haggard, 1989; Strain, Patterson, & Seidenberg, 1995), eye fixation times in reading (Inhoff & Rayner, 1986; Just & Carpenter, 1980; Rayner & Raney, 1996; Schilling, Rayner, & Chumbley, 1998), semantic categorization (Forster & Shen, 1996; Monsell et al., 1989) and lexical decision (Allen, McNeal, & Kvak, 1992; Forster & Chambers, 1973;

Monsell et al., 1989). Several studies have shown a word frequency effect in a variety of languages (Brysbaert, Lange, & Van Wijnendaele, 2000; Hino, Lupker, Ogawa, & Sears, 2003), while others have explored its interactions with other variables (Barry, Morrison, & Ellis, 1997; Reynolds & Besner, 2004). There has also been research into the neural correlates of word frequency effects (Voyer, 2003), as well as differences in event-related brain potentials for high- and low-frequency words (Rugg, 1990; Sereno, Rayner, & Posner, 1998; Van Petten & Kutas, 1992).

1.2.1. The Nature and Locus of Word Frequency Effects.

Most models of visual word recognition account for word frequency effects in one way or another, and in fact, many of those models have been explicitly implemented to account for word frequency effects. Although different approaches can be taken to classify models of visual word recognition (e.g., Carr & Pollatsek, 1985; Jacobs & Grainger, 1994), Monsell (1991) provides a simple classification that is largely based on the way in which models accounts for word frequency effects; this classification is quite convenient for illustrating the current knowledge on this area. Monsell divides models of word recognition into four broad categories: serial search models, localist activation models, hybrid activation-verification models and connectionist models. The following does not attempt to be an exhaustive description of all models of visual word recognition, but to provide an illustration of how different models account for frequency effects.

Serial search models

Most serial search models are inspired by the work of H. Rubenstein (Rubenstein, Garfield, & Millikan, 1970; Rubenstein, Lewis, & Rubenstein, 1971, 1971) and were originally developed to explain experimental results obtained using the lexical decision task. In more recent years, the chief proponent of this type of model has been K. Forster and his colleagues (Bradley & Forster, 1987; Forster, 1976, 1979, 1989, 1990, 1992; Forster & Bednall, 1976; Murray & Forster, 2004).

In this family of models, word recognition is analogous to searching for a book in a set of index cards in a library (Harley, 2001). All words known to a person are represented in a peripheral "access file" where entries are organized in decreasing order of frequency. The entries in this access file function as memory addresses that point to the complete word entries in the lexicon (or "master file"), where information such as meaning, pronunciation, part of speech and spelling is kept. When a target input is being considered for recognition, it is first transformed into an abstract orthographic representation that is compared sequentially with each entry in the access file; if a match is found before the end of the list, the target word is recognized and its representation in memory is accessed. High frequency words are recognised faster than lower frequency ones because they are compared to the target word's representation in the access file earlier during the search process.

Comparing each input letter-string with every single word in the lexicon is clearly not a feasible way to explain word recognition, so in Forster's (1976) model the access file is subdivided into "bins" that contain groups of words that share some characteristics, such as their first syllable or their first few letters. Within each bin, words are organized in decreasing frequency order. When a target letter-string is presented, the appropriate bin is identified according to the orthography of the words and the search for a matching pattern occurs only within that bin. Once again, high frequency words are compared with the target first, and therefore are recognised faster than low frequency words. In the lexical decision task, a "non-word" response would be returned if the search reaches the end of the respective bin without finding a match for the input letter string. More recent versions of this model

(Forster, 1989, 1992; Murray & Forster, 2004) have put forward the idea of a parallel search, in which all bins are searched simultaneously during the recognition process.

A central point in Forster's serial search model is that access files are modality specific: there is a separate access file for orthographic representations (for visual recognition), one for phonological representations (for auditory recognition) and one for syntactic-semantic information. Serial search models are therefore said to be *autonomous*, since there is no sharing of information between levels. The order in which words within each bin are compared with the input only depends on the frequency of exposure to those words in the relevant modality; in the case of visual word recognition, only written frequency would have an effect on the search order within each bin. The mental lexicon or master file is modality independent, allowing for some context effects in the shape of associative priming within the master file.

Localist Activation Models

The defining factor for this family of models is the idea that perceptual information from the input to the model (i.e., the word to be read) is directly fed into simple "feature detectors". This mechanism has taken different names in the literature, for example, Treisman (1964) called them "dictionary units", and Selfridge and Neisser (1960) called them "cognitive demons", while Morton (1969) called them "logogens". In fact, Morton's logogen model (Morton, 1969, 1979, 1980), along with the interactive activation model (Glushko, 1979; McClelland, 1987) are the two of the most influential incarnations of word-detector models.

In the logogen model, each feature detector has its own resting level of activation. During the word recognition process, perceptual information about the target input is gathered and passed on in parallel directly to all the logogens in the system. The level of activation of each detector unit increases if it shares features with the target input; for example, if the input contains the letter "*a*", the activation level of all logogens that contain that letter is raised a bit. A word is recognized when evidence for the identity of the word, represented by the activation level of one of the logogens, reaches a certain acceptance threshold and "fires", allowing access to the word's semantic and syntactic properties. The appropriate entry in the lexicon will normally reach the recognition threshold first because it receives activation from all the incoming sensory features of the input.

Resting levels are individually determined for each logogen and can be modified in two different ways. Firstly, context information can provisionally increase the resting level of some logogens, which accounts for context facilitation effects such as semantic priming. Secondly, every time that a word is recognized, the resting level of its logogen is increased a bit in a more or less permanent manner. Effectively, this lowers the recognition threshold for that word, so that less perceptual information is required to accumulate enough activation to trigger a response. High frequency words are recognized more quickly than low frequency ones because the more times a word has been encountered, the less is the distance between its resting activation level and its recognition threshold.

In the original version of the logogen model (Morton, 1969), each word had a unique detector unit that processed both visual and auditory inputs, so that exposures to a word in any modality had the same effect of lowering its detection threshold. However, more recent versions of the model have introduced separate visual and auditory inputs in order to accommodate evidence of differing effects of stimuli presented in each modality, such as the absence of cross-modal priming (Morton, 1979; Winnick & Daniel, 1970).

Despite the fact that this type of model was initially developed to account for findings from the word identification task, Coltheart, Davelaar, Jonasson and Bessner (1977) have proposed an updated version of the logogen model that was able to account for results from the lexical decision task. In this model, non-word responses are produced when none of the logogens has reached the detection threshold before a deadline that is dependent on the total activation of the entire system.

Some of the main principles behind the logogen model have been implemented in early connectionist models such as the "interactive activation" models of McClelland and Rumelhart (1981), Rumelhart and McClelland (1982) and others (e.g., Glushko, 1979; McClelland, 1987). These models were originally developed to account for the word superiority effects (that is, that letters are easier to identify while in the context of a real word), but they are usually considered general models of lexical access as well. McClelland and Rumelhart's interactive activation model consists of a localist network composed of detector units organized in three basic processing levels: a set of visual feature detectors at the input level, an intermediate level with letter representations and an output level where whole words are represented. Feature detector units correspond to very basic visual shapes that can be combined to form letters, for example, the letter "N" can be decomposed into two vertical lines joined by a diagonal line descending from left to right. Letter units are position-dependent, so that there are different representations (units) for the letter "a" as the initial letter of a word and for the letter "a" as the second letter of a word. Each unit in the model is connected to units on adjacent levels by either excitatory or inhibitory links, depending on whether they share features or not. For example, a unit at the letter level that corresponds to the letter "*C*" in the first position will have an excitatory link with all the words that start with the letter "*C*" and will have an inhibitory link with all other words. An increase in the level of activation of any given unit would spread in such a way that it increases or decreases the level of activation of the units to which it is connected depending on whether the links between them are excitatory or inhibitory. Units within the letter level and within the word level are connected via inhibitory links. The weights of the connections between units are hard-wired into the system, that is, there is no implementation of a learning mechanism in this model.

In this type of model, processing occurs simultaneously at all levels, with bottom-up and top-down information interactively affecting the outcome of the word recognition process, and with all letters of the input pattern being processed in parallel. When a word is presented for recognition, activation flows from the feature level to the letter level, and then to the word level, which in turn reinforces the appropriate units at the letter level and inhibits irrelevant units. This process continues until the system settles down to a state of equilibrium in which the appropriate word remains with a high level of activation while the other words are inhibited.

In the interactive activation model, word frequency effects are accounted for by assigning different baseline levels of activation to each word unit according to a word's frequency, so that units that represent higher frequency words have a higher level of activation to start with.

One particular model of reading that has gained prominence in the word recognition literature is the dual route model of reading proposed by Coltheart and his colleagues (Coltheart, Curtis, Atkins, & Haller, 1993; Coltheart & Rastle, 1994; Coltheart et al., 2001; Rastle & Coltheart, 1998; Rastle & Coltheart, 1999, 1999). This model shares important features with the interactive activation model. In the "lexical" or "direct" route a word is read by matching it to an existing entry in the mental lexicon, which allows access to the word's pronunciation. The process by which this lexical entry is accessed is akin to the one described in the interactive-activation model (McClelland & Rumelhart, 1981). In the "non-lexical" or "indirect" route, the pronunciation of a letter-string is assembled using a set of grapheme-phoneme correspondence rules (GPC rules) that map individual graphemes into individual phonemes which then yield access to the word's lexical representation. When a letter-string is presented to the word recognition system, both routes start processing it at the same time, and word recognition performance is determined by whichever route finishes the process first. For skilled readers, the direct lexical route is more efficient than the non-lexical route so it generally determines the level of performance in reading. However, the lexical route is not able to deal with reading non-words, which do not have a corresponding representation in the lexicon, so in those cases performance is determined by the non-lexical route.

In dual route models (e.g., Coltheart et al., 2001), word identification is not necessarily mediated by phonology, since meaning can be accessed directly from orthography via the lexical route, and word frequency affects the baseline level of activation of the orthographic units in the lexical route, pretty much in the same way as in the interactive-activation models.

Activation-verification or Hybrid models

This category includes models that share some characteristics with each of the previous two types of models. There are some variations between these models, but they all share the same basic characteristics. In general, they divide the word recognition process in two separate stages: "activation" and "verification". During the activation stage, a word-detector mechanism generates a small set of candidates that roughly share some characteristics with the target letter-string via a process akin to the one proposed by word-detector models (i.e., a word is considered as a candidate in the second stage if its activation reaches a certain threshold). During the verification stage, more detailed information about these candidates is retrieved from the lexicon and is serially compared one at a time with the target letter-string. The order of this comparison process is dictated by word frequency, as in the serial search models, so that high frequency words are considered first and can therefore be identified quicker. Word frequency plays no role during the activation process.

Models of this type include Becker's verification model (1976; Becker, 1980); Paap, Newsome, McDonald and Schvaneveldt's activation-verification model (1982; Paap, McDonald, Schvaneveldt, & Noel, 1987); and Norris' post-access checking model (1986).

Distributed Connectionist Accounts

There is a wide variety of distributed connectionist models of visual word recognition that vary in their specific architecture and implementation on issues such as the number of units and layers that form the network, the learning algorithm used, the existence and nature of feedback loops, and other factors (e.g., Harm & Seidenberg, 1999; Plaut, McClelland, Seidenberg, & Patterson, 1996; e.g., Seidenberg & McClelland, 1989). In general, these models consist of neuron-like units or "nodes" connected to each other via links that carry different weights. The weights of these connections determine how much the activation level of one node will affect other nodes to which it is connected. A defining characteristic of distributed connectionist models is that there are no specific representations associated to each word (such as logogens, or word units in the interactive activation model), but knowledge of different words is represented through different patterns of activity over the same set of units; a word is represented by the sum of active units in each of the levels. Another important element is that these models are able to implement a learning mechanism through the modification of the weights in the connections between nodes. A representative example of distributed connectionist networks is Seidenberg and McClelland's (1989) model of reading aloud. Their network is composed of three layers of nodes: one set of orthographic input nodes, one set of intermediate, hidden units and one set of phonological output nodes. In a similar way to the interactive activation model, each node has an activation level and is connected to all the units in the other two layers via weighted connections. The level of activation of each unit is defined by the sum of all the inputs (positive or negative) that feed into that unit.

Seidenberg and McClelland's model does not assume any prior knowledge embedded in the network, but uses the back-propagation algorithm to "learn" the correct associations between outputs and inputs (in the case of this model of reading aloud, the correct associations between phonological and orthographic patterns). Back propagation is an iterative process that involves the modification of the weights between the nodes in order to reduce the difference between the actual output of the network and the desired output. During the learning phase, the network is presented with a set of learning input patterns that represent written words and are asked to generate a phonological output. Each of the inputs in the training set is associated with an ideal output (corresponding to their correct pronunciation) so that these pairs function as exemplars to the network. The input patterns are presented to the network and they spread activation through the network's nodes until a pattern is generated in the output units. At the end of each pass through the network, the actual outputs generated by the network are compared to the exemplar outputs. The error or difference between real and ideal outputs is then used to modify the weights in the connections between units so that in the next pass this difference is reduced. This process is repeated a number of times, and each pass improves the performance of the network in generating the correct output patterns for the input exemplars. It is considered that the network is "trained" when its outputs reach a pre-defined criterion of similarity with the ideal outputs. When the network has been trained, the network is not only able to generate correct outputs for the training set, but is able to generalize this knowledge to other input patterns with similar characteristics.

In this type of model, words that are presented more often to the network during the training phase have more opportunities to influence the weights between connections so that they generate a better mapping between input and output for that particular pattern set. Frequency effects are therefore attributed to the learning process that defines the network architecture in skilled performance, and not to processes related to the recognition stage itself, as in the other models discussed in this section.

To recapitulate, serial search models typically account for frequency effects in terms of the order in which words in the lexicon are compared to the input, logogen-type models by contrast explain word frequency effects in terms of lowered recognition thresholds (or higher activation baselines) for high frequency words, while connectionist accounts attribute frequency effects to differences in the quality of the mappings between inputs and outputs, which are dependent on the number of exposures to each pair during the training phase.

1.2.2. Measuring Word Frequency.

Estimates of word frequency are usually assembled by counting the number of occurrences of each word in a relevant sample from a given language. These counts are most often presented in a normalized scale, in terms of "counts per million", as a way of estimating the *proportion* of usage or exposure to a particular word in the context of that language.

Word frequency counts attempt to be representative of general language use and different criteria have been applied to reach this goal; in general this is achieved by using a variety of sources, or by increasing the size of the sample from the same source. For example, the Brown corpus, on which the Kucera and Francis (1967) norms are based, is divided in two main sections, one of "Informative Prose" and another of "Imaginative Prose", and includes materials taken from newspaper reportages, press editorials, memoirs, religion, science fiction, detective fiction, and romance novels.

Some frequency databases are based exclusively on written materials (Carroll, Davies, & Richman, 1971; Kucera & Francis, 1967; Zeno, Ivens, Millard, & Duvvuri, 1995), others only on spoken materials (Brown, 1984; Dahl, 1979; Howes, 1966) and some include both modalities to one extent or another (Aston & Burnard, 1998; Baayen, Piepenbrock, & Van Rijn, 1995). In general, databases for written frequency counts are far larger than for spoken frequency, mainly because collecting and processing spoken-language samples is difficult and time consuming. For example, the spoken portion of CELEX (Baayen et al., 1995) is based on one million tokens, while the written portion is based on about 16.9 million tokens. In the case of the British National Corpus, the two databases for spoken English comprise about 10 million tokens, versus nearly 90 million for the written database.

The earliest word frequency counts in English were collected by Horn (1926), French, Carter, and Koenig (1930), Thorndike (1921) and Thorndike and Lorge (1944). These frequency lists were based on samples of only a few thousand words, mainly due to the difficulties associated with counting and tagging the words on the corpora by hand. An important development came about with the publication of the Kucera and Francis (1967) norms, a word frequency list based on the one million tokens of the Brown corpus, by far the largest sample to that date. The Kucera and Francis norms were the dominant resource for assessing word frequency until quite recently and are still widely used in different areas of research (e.g., Lipinski & Gupta, 2005; Nam, Lee, & Lee, 2004; Ward & Maylor, 2005), even when larger databases are now available.

In more recent years, the advent of computers and electronic text has facilitated the creation of much larger word frequency norms and made them easier to use. Some of the frequency norms that are widely used nowadays (and that are relevant to the present dissertation) include CELEX (Baayen et al., 1995), the British National Corpus (Aston & Burnard, 1998), the Educator's Word Frequency guide (Zeno et al., 1995), and the Hyperspace Analogue to Language (HAL) frequency list (Lund & Burgess, 1996).

The CELEX database includes frequency counts for Dutch, English and German. The English portion of CELEX is based on the 1991 version of the COBUILD/Birmingham corpus and comprises about 17.9 million tokens, with about 16.6 millions of them coming from written English and the rest from spoken sources. The written portion of the corpus was drawn from a set of 284 contemporary written texts that includes newspapers and books, while the spoken portion from sources such as the BBC World Service and telephone conversations. CELEX includes one list of frequencies for wordforms and one for lemmas. In the wordform list, frequency is calculated based on the individual types (for example, *child* and *children* have separate entries in this database). In the lemma list, the meanings of words were disambiguated and only word roots are presented. The frequency count in this database corresponds to the sum of all related words (i.e., the frequency count for the lemma "child" would correspond to the sum of frequencies of *child* and *children* wordforms). For experiments involving single word recognition, the relevant variable is usually wordform frequency, as there is no opportunity to disambiguate meaning as it was done for the lemma frequency list. CELEX is nowadays one of the most used word frequency databases, with at least 500 citations to date.

The British National Corpus is divided in three main word frequency databases. The largest one of them presents frequency counts for written materials and contains 89.7 million tokens and 921,074 types. The corpus from which these norms were compiled is comprised of texts from 3261 written sources drawn mainly from books and periodical publications (i.e., newspapers, journals and magazines), but it also includes published and unpublished ephemeral materials such as letters, essays, internal office documents, leaflets and others. These materials cover a wide variety of subjects such as natural, social and applied science, commerce, arts, belief and thought, leisure, world affairs and creative writing. The vast majority of texts included were published after 1975, making the content of the corpus more contemporary. The other two databases, the "demographic database" and the "context-governed database" are dedicated to spoken frequency counts.

The demographic (or "demog") frequency database contains 4.2 million tokens and 54,652 types. The frequency counts were drawn from spontaneous conversational material recorded from a representative sample of volunteers from 38 regions across the United Kingdom. The sample included male and female volunteers from different ages and social groups. The collection process was carried out by asking volunteers to unobtrusively record their conversations over two or three days. The context-governed (or "cg") part of the spoken corpus comprises 6.2 million tokens and 79,906 types and is drawn from four broad categories of social context. These contexts include educational and informative events (e.g. lectures, news broadcasts, tutorials, classroom discussions); business (sales demonstrations, meetings, interviews); institutional and public events (sermons, political speeches, parliamentary proceedings); and leisure events (e.g. sports commentaries, after-dinner speeches, club meetings). The materials in this database are more formal or scripted than the spontaneous conversations in the demog database.

The Educator's Word Frequency Guide provides frequency counts based on reading materials intended for children in the different grades of the American educational system. The corpus on which the frequency counts are based was gathered from approximately 6.300 books that include textbooks, works of literature, and popular works of fiction and nonfiction. It is organized according to school grade levels, starting with reading materials intended for children in first grade and progressing all the way to 12th grade, with an additional category (named "13+") for reading materials intended for people with post-secondary education.

The Hyperspace Analogue to Language (HAL) frequency list was developed as the language input for a computational model of human memory (Lund & Burgess, 1996). This corpus consists of nearly 131 million tokens collected from about 3,000 Usenet newsgroups during the month of February, 1995. According to Burgess and Livesay (1998), the size of the sample used, the conversational nature and the wide variety of topics covered in the postings used as source materials provide better estimates of word frequency than other frequency databases based on smaller corpora and more formal source materials. However, and equally due to the informal character of these materials, a relatively large proportion of the types listed in HAL's frequency list correspond to proper names, nicknames, misspellings and non-word symbols that introduce some noise into the frequency estimates.

1.2.3. Reliability and Validity of Word Frequency Counts

When considering the validity of word frequency counts, two prominent factors should be taken into account: the size of the language corpus used to estimate the frequency count, and its representativeness of language at large. In general, the statistical standard error of a frequency estimate correlates with the square root of the size of the sample corpus, so frequency counts based on larger corpora are usually considered to be more valid (Lee, 2003). However, the size of the sample is not the only factor; norms that attempt to estimate word frequency in the general language should include samples of all major types of text in proportions that are somehow related to their use (Clear, 1992; Leech, 1993). Frequency counts based on a single source tend to yield somewhat skewed frequency counts depending on the peculiarities of the chosen source. For example, the Marcus, Santorini, and Marcinkiewicz (1993) frequency norms are based on a very large corpus collected from the "Wall Street Journal", which leads to an over-representation of words such as *stock*, *margin* and *inflation*.

In order to compare frequency estimates derived from spoken and written corpora, Lee (2003) compared frequency counts for a random sample of 2807 words as estimated by three frequency lists from each modality. The three written frequency lists were the Educator's Word Frequency Guide (Zeno et al., 1995), the American Heritage Word Frequency Book (Carroll et al., 1971) and the Kucera and Francis (1967) word frequency norms. The three spoken frequency lists were the ones compiled by Dahl (1979), Brown (1984), and Howes (1966). Lee found that the correlation among written frequency estimates (mean $r = .79$) was higher than among spoken frequency estimates (mean $r = .66$).

Despite these apparently large correlations between databases (particularly for the written frequency counts), closer inspection reveals important differences between them. Burgess and Livesay (1998) compared the Kucera and Francis (1967) frequency estimates for 8,208 nouns with estimates from the much larger HAL (Lund & Burgess, 1996) frequency norms. They found that the overall correlation coefficient between the two databases for those words was .96, however, when the word sample was divided into three frequency bands, the authors found that the correlations for medium and low frequency words were only .12 and .14, respectively. Burgess and Livesay also found that the HAL frequency database provided much better predictions of reaction times on lexical decision than the Kucera and Francis norms on all frequency ranges. In general, frequency estimates calculated from smaller corpora tend to be somewhat unreliable, as these words are subject to a larger sampling bias, especially for low frequency words (Carroll, 1967, 1970).

Balota, Cortese, Sergent-Marshall, Spieler, and Yap (2004) also looked at how well different frequency databases can predict performance on different word recognition tasks. They found that there were differences of as much as 10% in the amount of variance that each database could account for. Kucera and Francis (1967), the smallest of the databases considered, accounted for the least variance, while the larger databases such as the Educator's Word Frequency Guide (Zeno et al., 1995) were much better predictors of performance.

Zevin and Seidenberg (2002) propose that, given the variability observed among different frequency estimates and the possibility that large sampling errors in the elaboration of frequency counts may generate spurious experimental results, a sensible strategy to follow would be to use several frequency databases when designing experiments where word frequency plays an important role.

1.3. Age of Acquisition Effects in Visual Word Recognition.

1.3.1. Background

Carroll and White (1973) were the first to propose that the age at which a word is first learned, the so-called Age-of-Acquisition or AoA, has a significant effect over lexical access performance so that words that are learned earlier in life are recognized faster in adulthood. They based this claim on two findings: a very high correlation ($r = .77$) between the AoA of picture names and picture naming latencies, and the results of a multiple regression analysis that yielded AoA as the single significant contributor to picture naming speeds, with no significant contribution from word frequency (Carroll & White, 1973, 1973). Carroll and White (1973) observed that "the finding that age of acquisition is more important than word frequency in a picture naming task raises serious questions concerning the interpretation of large numbers of verbal-learning, psycholinguistic, and reading studies in which frequency has been used as a critical variable. It is possible that age of acquisition is more relevant in such studies" (p.94).

Since this initial finding, a growing corpus of research has confirmed this advantage for early-acquired words in picture naming (Barry et al., 1997; Bonin, Chalard, Meot, & Fayol, 2002; Ellis & Morrison, 1998) and extended the findings to other tasks such as lexical decision (Brysbaert et al., 2000; Gerhand & Barry, 1999b; Morrison & Ellis, 1995, 2000), auditory lexical decision (Turner, Valentine, & Ellis, 1998), word association tasks and semantic categorization (Brysbaert, Van Wijnendaele, & De Deyne, 2000), eye fixation durations in sentence reading (Juhasz & Rayner, 2003), and speeded word naming (Gerhand & Barry, 1999a). Age of Acquisition effects have been found in a variety of languages including Japanese (Yamazaki, Ellis, Morrison, & Ralph, 1997), Chinese (Guan & Fang, 2002), Dutch (Brysbaert et al., 2000), Spanish (Cuetos & Alija, 2003), Italian (Colombo & Burani, 2002), and French (Bonin, Chalard, Meot, & Fayol, 2001). Other studies have found that AoA is a good predictor of language performance in patients with language impediments such as aphasia (Cuetos, Aguado, Izura, & Ellis, 2002; Ukita, Abe, & Yamada, 1999) and semantic dementia (Lambon-Ralph, Graham, Ellis, & Hodges, 1998; Taylor, 1998).

Despite the striking implications of these results, AoA was largely ignored in studies and models involving word recognition for many years. Most of the publications dealing with frequency cited in the previous section failed to control for AoA, and even more recent studies have also ignored the issue of AoA. For example, as noted by Bowers, Davis and Hanley (in press), none of the neighborhood studies they reviewed controlled for AoA, including many recent studies. Furthermore, most connectionist models of word recognition have not even considered AoA as a possible influencing factor on network performance and rely heavily on frequency effects for their representations (e.g., Patterson, Seidenberg, &

McClelland, 1989; Plaut, 1997; Plaut et al., 1996; Seidenberg & McClelland, 1989). In fact, some authors have even proposed that it is not possible for a connectionist network to show AoA effects due to the phenomenon of catastrophic interference in distributed models (Gerhand & Barry, 1998; Morrison & Ellis, 1995).

As mentioned before, not many models of word recognition include an account of AoA as they do with frequency, so most of the discussion on this topic has not evolved around the validity of theoretical models but around the more pragmatic question of the locus of the effect, which in turn allows for predictions to be tested on the models. Many hypotheses have been proposed to explain the fact that early-acquired words are better recognized than later-acquired words. The simplest potential explanation is that, in fact, there is no real AoA effect and that the observed advantage for early-acquired words is only a result of a confound between word frequency and AoA: earlier words are usually encountered more often through out life. This issue will be discussed further in Chapter 2. Among studies that accept the authenticity of the AoA effect, several explanatory hypotheses have been proposed.

The phonological completeness hypothesis and other phonological accounts.

Early studies showed evidence that AoA effects were more prominent in tasks that required access to phonology, leading some authors to propose the phonological output system as the locus of AoA effects. For example, Gilholy and Watson (1981) noted that AoA effects were larger for word-production than for word-recognition tasks, and attributed these effects to differences in the speed at which early and late words are retrieved from the phonological output lexicon. They proposed that AoA (and not word frequency) was the critical factor in determining the activation thresholds within the phonological output system in word-activation models. Levelt, Roelofs, and Meyer (1999) also suggested that AoA effects could arise because of differences in the speed with which lexemes can be accessed from lemmas during the lexicalisation process, while Gerhand and Barry (1998) found that it took longer to pronounce late-acquired words than early-acquired words.

Brown and Watson (1987) provide a theoretical framework to justify a phonological locus for AoA effects. They suggested that AoA effects arise because early-acquired words have less fragmented phonological representations than late-acquired words. Phonological representations of early-acquired words are stored as whole units, since there are no major storage constrain when few words have been acquired, but as vocabulary grows, so do demands on memory, and a more efficient storage strategy is adopted for later-acquired words: their phonological representations are divided up into smaller phonetic units (e.g., syllables or phonemes), and only minimal information about the word's pronunciation is stored. When an early word is accessed, its phonological representation is readily available, while later words must be re-generated from their fragmentary components. This assembly process, required for later-acquired words but not for early ones, would explain the extra time taken to recognize and produce late-acquired words. This "Phonological Completeness" hypothesis has been cited often in the literature as a possible explanation for AoA effects (Barry, Hirsh, Johnston, & Williams, 2001; Ellis & Morrison, 1998; Gerhand & Barry, 1998; Morrison & Ellis, 1995; Morrison, Ellis, & Quinlan, 1992).

Metsala and Walley (1998) proposed an alternative phonological account of AoA effects. In this account, the "Lexical Restructuring" hypothesis, AoA effects also emerge as a result of differences in the quality of phonological representations according to the age in which they were acquired, but the character of this advantage is opposite to the one proposed by Brown and Watson (1987). According to Metsala and Waley, all words undergo a restructuring process by which their phonological representations are gradually segmented down to the phoneme level. This restructuring process is advantageous to the recognition and production of words, so that words that achieve a more complete segmental restructuring are processed more efficiently. Early-acquired words undergo this restructuring process earlier, and achieve more fine-grained phonological representations (and hence better performance) than late-acquired words.

Although phonological accounts of AoA effects have enjoyed some popularity in the literature (particularly the phonological completeness hypothesis), their validity has come into question on the basis of more recent findings. Much of Brown and Watson's (1987) claim was based on the assumption that AoA effects could only be observed in tasks that involved phonological coding, such as word or object naming, but AoA effects have also been found in tasks for which access to phonology is not required. There are many studies showing AoA effects in lexical decision (e.g., Brysbaert et al., 2000; Gerhand & Barry, 1999b; Morrison & Ellis, 1995, 2000), a task for which phonology is not necessarily activated. Even if one took the view that there is an involvement of phonology in lexical decision (e.g., Pexman, Lupker, & Jared, 2001), AoA effects have been founds in tasks for which there is no ovbious way in which phonology would need to be accessed. For example, Vitkovitch and Tyrrell (1995) found AoA effects when they asked participants to distinguish between real and imaginary objects. Age of acquisition effects were also found by Brysbaert, Van Wijnendaele, and De Deyne (2000) in a semantic categorization task in which participants had to decide whether the stimuli presented were nouns with a definable meaning or first names. More recently, Ghyselinck, Custers and Brysbaert (2004) found an AoA effect using the semantic Simon paradigm, a task in which phonology does not necessarily take part.

Monaghan and Ellis (2002) used a word segmentation task to test whether AoA effects have a phonological locus. In this task, participants are asked to remove the initial parts of a word (as indicated by a cue) and pronounce what remained of the word as quickly as possible. The phonological completeness hypothesis predicts that it should take longer to segment early-acquired words since they are stored whole, while late acquired words (which are already stored as segments) should be faster to segment. On the other hand, the lexical restructuring hypothesis predicts that the segmentation of early-acquired words should be faster because their representations are better segmented than late-acquired words. Monaghan and Ellis also point out that the overall quality of phonological representations in an individual is dependent on the general phonological skills of that person. This led them to speculate that if AoA effects are a result of the quality of phonological representations of words (as predicted by phonological hypotheses), there should be a strong relationship between the phonological skill of a participant (as measured by the word segmentation task) and the size of the AoA effect shown by that participant. If a participant showed a high level of phonological ability, one would expect larger AoA effects. In Monaghan and Ellis' study, there was no advantage for late acquired words in any of the conditions, which seems to contradict the idea that early words have more holistic phonological representations, as the phonological completeness hypothesis would predict. Also, in all except one condition, there was no advantage for early-acquired words, which seems to contradict the prediction derived from the lexical restructuring hypothesis. Additionally, the authors did not find a significant correlation between phonological skill and AoA effect size, leading them to conclude that AoA effects do not depend on the quality of phonological representations of words.

In view of this accumulation of evidence, it is difficult to maintain that the phonological output lexicon could be the sole locus of AoA effects. Most of the studies mentioned above do not explicitly rule out the involvement of phonology in AoA effect, but the evidence would, at most, allow for a distributed locus including the phonological output lexicon, as in the case of Ellis and Lambon-Ralph's (2000) account of AoA effects (see below for more detail on this possibility) .

The Growing Network Model and other semantic accounts.

As mentioned in the previous section, one of the main criticisms for phonological accounts of AoA effects is the presence of AoA effects in tasks that do not require phonology; many of such tasks are semantic in nature. There is indeed a growing body of research that could be interpreted as evidence for a semantic locus of AoA. For example, Van Loon-Vervoorn (1989) found an advantage for early-acquired words in word-association; in this task, participants were asked to name the first word that came to their minds when presented with a stimulus word. These results lead Van Loon-Vervoorn to propose that AoA is a semantic rather than a lexical variable. At the centre of this view is the idea that the organization of information in the brain is dependent on the order in which concepts are entered into the semantic system: meanings are tightly interconnected and newer concepts are defined in terms of what is already known (Brysbaert et al., 2000; Brysbaert et al., 2000; Van Loon-Vervoorn, 1989).

Brysbaert et al. (2000) highlight a series of findings that point towards semantic involvement on AoA effects. Among these findings, they note that several studies have shown a stronger correlation between AoA and other semantic variables (such as imageability, concreteness and number of meanings) than between AoA and frequency, as is the case with the results obtained by Rubin (1980), Whaley (1978) and Morrison, Chappell and Ellis (1997). Brysbaert et al. also point out that strong AoA effects have been found in object naming (e.g., Barry et al., 1997; Bonin et al., 2002; Ellis & Morrison, 1998), a task that has been shown to require semantic access in order to link the picture of the object with its name (e.g., Snodgrass, 1984; Theios & Amrhein, 1989).

More recently, Brysbaert et al. (2000) replicated Van Loon-Vervoorn's (1989) AoA effect in word association and also found a robust AoA effect in a semantic categorization task in which participants were asked to decide if stimuli presented to them were nouns or first names. Ghyselinck, Custers and Brysbaert (2004) also found an AoA effect using a variation of the semantic Simon paradigm. In this experiment, participants were presented with Dutch words that corresponded to either living (e.g., *dog*) or non-living (e.g., *sword*) entities and were presented sometimes in uppercase and sometimes in lowercase letters. One set of the words used was classified as "early acquired" and another as "late acquired". Participants were instructed to quickly classify each item according to the letter case, while disregarding the semantic category of the item itself. If an item was presented in uppercase letters, participants had to say "living", while if it was presented in lowercase letters, they had to say "non-living". Ghyselinck et al. replicated previous results in this task (De Houwer, 1998) that show that "congruent" responses are performed quicker; that is, participants were quicker to say "living" to words like *dog* than to words like *sword*, even though they were explicitly instructed to base their decision solely on the surface characteristics of the items and not their meaning. This indicates that the meaning of the stimuli were automatically accessed and interfered with the verbal response of participants. More interestingly, the authors found that this "semantic Simon" effect was significantly stronger (in fact, twice as strong) for early-acquired words than for late acquired words. Ghyselinck and her colleagues interpret this as evidence that the AoA effect in word processing tasks is, at least partially, due to semantic activation.

Steyvers and Tenenbaum (2005), in their Growing Network Model, provide a theoretical framework that can accommodate the claim that AoA effects are (at least partially) semantic in nature. This model proposes that AoA effects reside at the semantic level and result from a higher connectivity of early-acquired words as a direct consequence of the way in which semantic networks grow. Using graph theory, Steyvers and Tenenbaum analyzed the large-scale structure of three types of semantic networks: one based on the word free-association norms collected by Nelson, McEvoy and Schreiber (2004), another based on Roget's Thesaurus (1911 edition) and a last one based on WordNet, a database that organizes English words into synonym sets linked by different types of relationships (Fellbaum, 1998; Miller, 1995). Despite important differences in the composition of these semantic networks, the authors found that all of them exhibited a *small-world* structure (Milgram, 1967; Watts & Strogatz, 1998) which is characterized by sparse connectivity, short average path-lengths between words, and strong local clustering. In this type of network, a relatively small number of nodes are very well connected to other nodes, acting as "hubs" in the connections of the network and the distribution of the number of connections for a given node follows a power function (with a few very well connected nodes and many less well-connected ones).

These characteristics place strong constraints on the way in which networks such as these can be generated and, in the case of language development, define how the semantic system is organized. One way of achieving this small-world structure is for the network's growth to be incremental and to follow a preferential-attachment principle (Barabasi & Albert, 1999). Under these conditions, when new nodes (i.e., concepts or words in our case) are gradually added to the network, they are more likely to be connected to old nodes that already have many connections, and not too likely to be connected with less well-connected nodes. The result of this process is that well-connected nodes become even better connected, and nodes that do not have many connections tend to stay that way. According to Steyvers and Tenenbaum (2005), "this growth process can be viewed as a kind of semantic differentiation, in which new concepts correspond to more specific variations on existing concepts and highly complex concepts (those with many connections) are more likely to be differentiated than simpler ones" (p. 5). This model predicts that words learned earlier in life will be more central to the network; that is, they will be better inter-connected than later-acquired words because later-acquired words attach to them, thus explaining the Age of Acquisition effect.

An important difference between this account of AoA and others discussed in this chapter is that early-acquired words do not have qualitatively better representations that late-acquired words, they are just more central to the semantic network and are better connected with other concepts.

Steyvers and Tenenbaum (2005) studied the relationship between the number of connections of each word in the three networks described before and two different measures of AoA (subjective ratings from Gilhooly and Logie, 1980a and objective norms from Morrison, Chappell and Ellis, 1997). They found that early-acquired words were better connected than late-acquried words, and that there was an interaction with word frequency so that AoA effects on connectivity are stronger for high frequency words. Steyvers and Tenenbaum also analyzed the correlation between AoA, word frequency and degree of centrality with respect to latencies for word naming and lexical decision. They found that, besides correlations for AoA and frequency encountered in other studies, centrality also correlated negatively with reaction times on both tasks; words that were more central in the semantic network were identified quicker. Combined, these findings seem to confirm the idea that AoA effects can be attributed to the advantage in connectivity conferred to early-acquired words in a growing semantic network.

However, Izura and Ellis (2002, 2004) conducted a series of studies on the effects of AoA in first and second language using translation pairs (words with the same meaning in two different languages). They posit that if, as evidence seems to show (Brysbaert, 1998; Costa, Miozzo, & Caramazza, 1999; de Bot, 1992), there is only one semantic system for words in both L1 and L2, translation pairs should share the same abstract semantic representation, and words in L2 should "inherit" the semantic characteristics of their equivalents in L1 (including AoA). If AoA had a semantic locus, performance on both languages should respond to the time in which the concept first entered the lexicon, independently of which language it belongs to, so that in this case AoA in the first language (L1) should be the better predictor of latencies even if the task is performed in the second language (L2). They found that lexical decision speed in L2 was not determined by the age at which words were first learned in L1, but by their AoA in L2. It is difficult to accommodate these results in the light of the growing network model and other semantic accounts of AoA (as long as one accepts the caveat that L1 and L2 share the same lexicon).

Distributed Connectionist Accounts.

Contrary to frequency effects, which lay at the center of most connectionist models of word recognition (e.g., McClelland & Rumelhart, 1981; e.g., McClelland & Rumelhart, 1985; Patterson et al., 1989; Plaut, 1997; Plaut et al., 1996; Seidenberg & McClelland, 1989), AoA effects have sometimes been thought to be incompatible with distributed processing networks that use back propagation as their learning algorithm (e.g., Gerhand & Barry, 1998; Moore & Valentine, 1998; Morrison & Ellis, 1995). Those accounts point out that in many distributed back-propagation connectionist models, later-acquired information tends to overwrite what was previously learned by the network in what has been typified "catastrophic interference" (French, 1999; McCloskey & Cohen, 1989) or the "stability-plasticity dilemma" (Grossberg, 1976). In back-propagation, the weights between nodes in the network are successively re-adjusted so that output patterns approach the exemplars provided to the model during the training phase. Very often in such simulations, learning patterns are entered sequentially in blocks, so that a set of patterns is used to train the network until a given performance criterion is reached. When a new set of patterns is then introduced, the weights in the network re-adjust to fit the new set of items, and performance on the first set drops suddenly and radically.

However, as Ellis and Lambon-Ralph (2000) point out, vocabulary acquisition does not usually happen in a sequential, blocked fashion. Instead, word learning is cumulative, with a mixture of training for earlier acquired words along new ones. Under those more natural circumstances of interleaved training, Ellis and Lambon-Ralph (2000) and Smith, Cottrell and Anderson (2001) propose that AoA effects are indeed intrinsic to distributed connectionist networks. According to these authors, the origin of AoA effects lies in a decrease of plasticity in the network. As the network adjusts its weights to accommodate early patterns, they become entrenched in the architecture of the model, to the detriment of latter acquired patterns; as Ellis and Lambon-Ralph put it "the network becomes increasingly committed to representing [early patterns] and, as a result, less and less able to assimilate new, late patterns" (p. 1107)

This loss of plasticity can be attributed to the way in which the learning algorithm affects the weights between nodes during training. In an untrained network, the weights between connections are initially set to random values between 0 and 1. Most of those weights tend to be grouped around the middle point of the range (the .5 value). The learning function for back-propagation networks is such that it has a larger effect on weights that are near the .5 value, while weights that are already at the extremes of the scale (0 or 1) are less affected. As the network reaches skilled performance levels for an early set of items, weight values tend to migrate to the extremes of the scale, so that when a new set of items is introduced, the learning algorithm cannot modify the weights in the network to the same extent that it could when the network was untrained (and the weight values tended to be grouped around .5). As a result of this, the quality of the input-output mappings for late acquired patterns is not as good as with earlier patterns. In this account, the AoA effects are located in the weights of connections between representational levels, rather than the representations themselves.

Originally, it was thought that the advantage for early-acquired patterns were independent of the model's characteristics, as long as interleaved training is used. Ellis and Lambon-Ralph (2000) and Smith, Cottrell and Anderson (2001) showed that AoA effects were present in variety of simulations in which they varied different aspects of the model such as network topology, learning algorithm, learning rate, input frequency, pattern frequency trajectory, vocabulary size, and performance criteria. The advantage of earlier patterns remained even when controlling the cumulative frequency of presentation of the patterns.

The patterns used by Ellis and Lambon-Ralph (2000) in their simulations were random binary strings of 100-bits with each bit having a probability of .2 of being set to one and the rest of the bits set to zero. Output patterns were generated by flipping bits from the input pattern from one to zero and vice-versa with a probability of .1; this generated a degree of regularity between input and output patterns. Zevin and Seidenberg (2002), however, pointed out that the patterns used in the simulations conducted by Ellis and Lambon-Ralph (2000) and Smith, Cottrell and Anderson (2001) were fundamentally different from natural language. In those two simulation studies, input and output patterns were randomly generated binary strings, and so they lacked the structure that in a natural language allows for the generalization of knowledge from one pattern to another. Indeed, when Zevin and Seidenberg used more realistic patterns (i.e. words) in their own simulations, they only found AoA effects when there was little or no overlap between patterns. The authors concluded that, due to the regularities found in real languages, any advantage obtained by early items during training would fade away in skilled performance because knowledge of early items would also help in the processing of later items. For example, the orthographic forms *prince* and *print* map onto similar output patterns (e.g., the phonological forms /prIns/ and /prInt/, respectively). Under these conditions, early learning (*prince* → /prIns/) contributes to late learning (*print* → /prInt/), eliminating any differences generated by AoA.

The Cumulative Frequency Hypothesis.

The cumulative frequency hypothesis proposes that the determinant variable in word (or picture) recognition performance is the total number of times that a word has been encountered, or as Lewis (1999) puts it, the accumulation of instances of a given stimulus, not AoA or surface frequency. Simply said, the more a stimulus has been encountered through life, the easier and faster it will be to retrieve its representations from memory. According to this account, both word frequency and the time of residence in a person's memory contribute to performance with equal weight: it is irrelevant whether a word has been encountered many times over a short period of time or just a few times per year over many years, what really matters is the total number of exposures to it. Proponents of this view (e.g., Lewis, 1999a, 1999b; Lewis, Gerhand, & Ellis, 2001; Zevin & Seidenberg, 2002), point out that this hypothesis parsimoniously accounts for two of the major factors (AoA and frequency) identified in the literature as determinants of performance in word recognition.

Carroll and White (1973) proposed that if cumulative frequency was the critical factor, some multiplicative combination of AoA and frequency would be a better predictor of word naming times in a multiple regression analysis than AoA by itself. They further speculated that if both AoA and frequency effects respond to a common mechanism (i.e., cumulative frequency), there should be a statistical interaction between these two variables so that the effect would be larger for early-acquired words than for late acquired words. However, Carroll and White did not find this to be the case, and most studies that have shown both AoA and frequency effects did not yield a significant interaction between the two factors (e.g., Gerhand & Barry, 1998; Ghyselinck et al., 2004). This additive character of the two variables was usually interpreted as evidence for separate loci for the AoA and frequency effects and against the cumulative frequency hypothesis. For example, Gerhand and Barry use this argument as supporting evidence for their view that, in word naming tasks, frequency affects word recognition while AoA affects phonological output. However, Lewis and colleagues (1999a, 1999b; Lewis et al., 2001) proposed a revised version of the cumulative frequency hypothesis that addresses the concerns raised by these findings. They point out that the additive character of AoA and frequency effects only points to separate loci if one assumes a linear relationship between cumulative frequency and word recognition performance. According to different theories of learning and the effects of practice, performance can be better represented by compressed functions of frequency, following either a power law (e.g., Newell & Rosenbloom, 1981) or an exponential law (e.g., Heathcote, Brown, & Mewhort, 2000) depending on which theory one adheres to. McCusker (1977) compared how well the logarithmic function, the power function and the exponential function could account for word naming performance, and concluded that an inverse logarithmic function provided the best fit for the data. Lewis et al. (2001) showed that when this transformation is used, the cumulative frequency hypothesis actually predicts the additivity of AoA and frequency effects. Furthermore, Lewis re-analyzed four previous studies (Balota, Cortese, & Pilotti, 1999; Carroll & White, 1973; Gerhand & Barry, 1998; Spieler & Balota, 1997) and found that when the logarithmic transformation was applied they were actually consistent with the cumulative frequency hypothesis instead of with independent AoA and frequency effects.

The cumulative frequency hypothesis predicts that the advantage of early-acquired words would tend to diminish as the age of the person increases. Take two words with similar surface frequencies but learned 10 years apart from each other. At an early age, the cumulative frequency of each of these words would be considerably different, but the absolute size of this difference would remain fixed as the person grows older, while its relative size would decrease. Take for example the words *bubble* (AoA = 4.26 years) and *organic* (AoA = 12.92 years), which have the same CELEX surface frequency (F = 15). For a 16 year old person, the ratio of cumulative frequency estimates between the two words would be about 3.8, while at age 60 this same ratio would be only 1.2. However, the difference in AoA between the two words would always be 8.66 years (or a ratio of 3.03). If cumulative frequency were the relevant factor, the effect should decrease. Morrison, Hirsh, Chappel and Ellis (2002) tested this possibility and found that AoA effects for older participants was at least as strong as for young participants in both word and picture naming. This lead them to conclude that AoA effects are due to the age at which words are learned, or the order in which they are learned, but not to the time they are known (or cumulative frequency).

Ghyselinck, Lewis and Brysbaert (2004) evaluated the cumulative frequency hypothesis by systematically comparing the size of AoA and frequency effects for the same set of items in a variety of visual word recognition tasks. The tasks they used were: tachistoscopic perceptual identification, three variants of word naming (immediate, delayed and speeded naming), three variants of lexical decision (with illegal non-words, legal non-words and pseudo-homophones) and a semantic categorization task in which participants had to decide whether the stimuli presented were nouns or proper names. Each of these tasks is believed to rely to a varying extent on different types of information, from phonology to semantics, and the authors proposed that differences in the size of frequency and AoA effects between the tasks would help determine the locus of such effects. For example, they argued that if AoA effects were the result of the organization of the semantic system, one should expect them to be present in the semantic classification task but not necessarily in the perceptual identification task. Similarly, if AoA effects were only derived from the way in which the phonological output lexicon is organized, the effects should be more prominent in the word naming tasks and the lexical decision task with legal nonwords than all the other tasks. Additionally, Ghyselinck et al. were interested in determining the degree to which AoA and frequency effects correlate to each other. They argued that if both effects had a common origin, as posited by the cumulative frequency hypothesis (e.g., Lewis et al., 2001) and some distributed connectionist accounts (Ellis & Lambon-Ralph, 2000; Smith, Cottrell, & Anderson, 2001), there should be a strong correlation between the two variables, but this would not necessarily be the case if frequency affected lexical perception processes while AoA affected output processes, as proposed by some authors (e.g., Gerhand & Barry, 1998). They also pointed out that the cumulative frequency hypothesis predicted similar weights for AoA and frequency when the frequency and time known factors were log-transformed, as proposed by Lewis (2001). As summarized in the following sequence of equations, this prediction is a direct result of the logarithmic transformation of cumulative frequency:

Cumulative frequency = surface frequency * time of residence

Time of residence = Age - AoA

RT = $a + b$ * log (cumulative frequency)

RT = $a + b$ * log (surface frequency * time of residence)

RT = $a + b$ * log(surface frequency) + b* log(time of residence)

RT = $a + b$ * log(surface frequency) + b* log(Age-AoA)

(where a and b are free parameters of the model)

As can be seen, the multiplying factor b is the same for AoA and frequency, so one would expect both variables to carry a similar weight in performance.

To Summarize the results, Ghyselinck et al. (2004) found that both AoA and frequency effect were present in tasks that tap into different stages of the word recognition process, thus pointing towards a broad origin of the effects and away from a specific stage of the word recognition process. There was a strong correlation between the two variables, but, contrary to what the cumulative frequency hypothesis predicts (see above), the weight of the AoA factor always seemed to be an order of magnitude larger than the weight of the frequency factor. The authors interpret these results as problematic for the cumulative frequency account and supportive of connectionist models such as the ones proposed by Ellis and Lambon-Ralph (2000) or the semantic organization accounts defended by Steyvers and Tenenbaum (2005).

1.3.2. Measuring AoA.

As with many factors that are dependent on the personal experiences of each individual, it is not possible to know the exact time in which a word has been learned by any given person, so authors have had to draw on estimates that reflect the actual AoA of words as closely as possible. Broadly speaking, there are three methods that have been used to estimate the age at which words are learned: (i) estimates based on subjective adult ratings of AoA (Bird et al., 2001; Gilhooly & Logie, 1980a); (ii) objective estimates derived from direct tests of word knowledge on people of different ages (Gilhooly & Gilhooly, 1980; Morrison et al., 1997) or on reports obtained from teachers or parents about children's abilities (Dale & Fenson, 1996; Kohnstamm, Schaerlaekens, de Vries, Akkerhuis, & Froonincksx, 1981); and (iii) estimates derived from grade-based word-frequency corpora (Carroll et al., 1971; Lete, Sprenger-Charolles, & Cole, 2004; Zeno et al., 1995). An extensive list of AoA norming studies can be found in Tables 2.8 and 2.9 of this thesis.

Adult ratings are probably the most used method in the field to estimate AoA. For these ratings, participants are usually presented with a randomized list of words and are asked to indicate approximately how old they were when they learned each of them. Most times, participants are instructed to specify when was the earliest age at which they could have understood, said or read each word, whichever came first (e.g., Gilhooly & Logie, 1980a), but on a few occasions, researchers are interested in the age of acquisition of a word for a specific modality (i.e., spoken or written). In these cases, raters are explicitly asked to provide their ratings with this in mind (e.g., spoken but NOT written AoA in Yamazaki et al., 1997). In some studies, participants simply indicate their AoA estimates in years (Ghyselinck, Custers, & Brysbaert, 2003; Ghyselinck, De Moor, & Brysbaert, 2000), but most researchers have chosen to use scales associated with age-ranges. For example, Gilhooly and Hay (1977) used a 7-point scale with the first point in the scale corresponding to an AoA of 2 years or less and the last point to an AoA of 13 years or more, and intermediate points corresponding to two-year intervals. This is the scale that has been used by some of the most influential AoA norming studies in English such as Gilhooly and Logie (1980a), Bird, Franklin and Howard (2001) and Morrison, Chappell and Ellis (1997). Other studies have chosen to use different size intervals for their rating scales, varying from 5 to 9 point scales (Alario & Ferrand, 1999; Colombo & Burani, 2002; Cuetos, Ellis, & Alvarez, 1999). Despite these slight differences, the general procedure followed for all these AoA ratings is quite similar, and there seems to be a high level of agreement between studies even when different scales were used; for example, ratings obtained by Carroll and White (1973), using the 9-point scale and ratings obtained by Carroll and White (1973) using an 8-point scale were highly correlated with a coefficient of .96 (see below for more on the reliability of AoA ratings).

As mentioned before, AoA estimates obtained by "objective" methods usually rely on direct vocabulary tests on children of different ages (Morrison et al., 1997) or reports obtained from teachers or parents about children's abilities (e.g., Kohnstamm et al., 1981). Vocabulary tests can be performed by either asking children to define words or by asking them to label pictures. For example, Gilhooly and Gilhooly (1980a, Experiment 2) read aloud a series of words and asked children to say what each of them meant. The objective age of acquisition for each of these words was taken as the average age, in months, at which 50% of participants would know the word. Morrison, Chappell and Ellis (1997) used a different approach: they asked participants of different ages to name a series of pictures and the AoA of a particular item was defined as the earliest age at which at least 75% of the raters were able to produce the name of its picture. Because naming responses to pictures are relatively unambiguous, this method has some advantages over the one used by Gilhooly and Gilhooly, which requires some interpretation when evaluating whether a child provided an adequate definition of each item.

In some studies, vocabulary knowledge is not assessed by testing children directly, but is based on the reports of adults that have usual contact with them, such as their parents or teachers. In such cases, the adults are asked whether children of a determined age group would know a word or not. An example of this type of study is the one conducted by Kohnstamm, Schaerlaekens, de Vries, Akkerhuis, and Froonincksx (1981), in which a sample of Dutch teachers were asked to choose which words from a large list would be understood by their 6 year-old pupils. Also, Dale and Fenson (1996) assembled a set of month-by-month AoA norms based on the Communicative Development Inventories. They asked a large sample of parents with children from 8 to 16 months to report which words their own child could understand and produce (out of a list of 396 items), and parents with children from 16 to 30 months to report which words their own child could produce (out of a set of 680 items).

A third way in which AoA can be estimated is by using grade-based word frequency counts. Most word frequency databases, like CELEX (Baayen et al., 1995) or the BNC (Aston & Burnard, 1998), provide frequency counts based on adult corpora. However, some databases, such as the *Educator's Word Frequency Guide* (WFG; Zeno et al., 1995) and

MANULEX (Lete et al., 2004), present cross-sections of word frequencies taken from materials aimed at different age groups, usually organized by school grades. In this type of frequency database, frequencies are usually calculated based on representative samples of texts, such as school texts and story books, that are specifically aimed at the different age groups. Each frequency count thus obtained is associated with an age range; for example, the WFG presents frequency counts from first to 12th grades (corresponding to the American primary and secondary educational system) and beyond secondary education (a category dubbed "13+"), while MANULEX provides similar frequency counts for three broad categories: first grade (approximately 6 years old), second grade (7 years-old) and a combined category that includes frequencies from third to fifth grade (8 to 11 years old).

These age-dependent frequency counts have been used by some authors to provide estimates of AoA. For example, Carroll and White (1973, 1973) obtained what they call an "Age of Word: Objective Data" by using age-marked databases that provide estimates of frequency for words read (Dale, 1948; Dale & Eichholz, undated) and written (Rinsland, 1945) by children at different school grades. Carroll and White used a somewhat complicated scheme to classify words with a 10-point scale representing the relative order in which they were acquired. The position of each word in the scale was based on the first school grade for which a word had a relatively high frequency in the various frequency databases they used (i.e., Dale, 1948; Dale & Eichholz, undated; i.e., Rinsland, 1945). Zevin and Seidenberg (2002; 2004) used the *Educator's Word Frequency Guide* as the basis for their analyses on cumulative frequency, which, as explained before, includes an AoA component.

1.3.3. Validity and Reliability of AoA Ratings

As can be seen from the previous section, obtaining "objective" AoA estimates is an arduous and complicated process that is often impractical to use for large sets of words.

Consequently, most studies in the field have opted to use, in one way or another, subjective estimates as a representation of the age at which words are acquired. As with any subjective judgment of an objective property, it is important to establish the reliability and validity of the ratings thus obtained. Authors in the AoA literature have gone to great lengths to show that AoA ratings are consistent with each other (both between raters in the same study and between different studies) and that they indeed reflect a measure of the age at which words were acquired.

A variety of methods have been used to show the reliability of subjective AoA measurements. Estimates given by different groups of raters, both within the same study and across different studies, tend to agree with each other. A simple way of assessing this point is by correlating estimates obtained by different groups for the same items. In general, researchers are interested mostly in capturing the relative order of word acquisition, not the absolute AoA value, so even if there are some variations in the means and ranges of the ratings between groups, a high correlation is usually taken as a sign of reliability. This is the case when the gender of the raters is taken into account. Adult female raters tend to give slightly earlier estimates of AoA than their male counterparts. For example, Carroll and White (1973) found that the average AoA score given by female participants to their entire list of words was 4.64, while male participants gave them a score of 4.83 (in a 9-point scale). Along the same lines, Winters, Winter and Burger (1978) report a difference of about five months between the average AoA estimates given by male and female raters. However, many studies show that there is a high correlation between ratings obtained from both genders. For example, Carroll and White (1973), Winters, Winter and Burger (1978) and Gilhooly and Hay (1977) all report correlation coefficients of more than .93 between ratings provided by male and female raters. Inter-group reliability for AoA estimates within the same study has also been shown when participants are split in other ways than gender. Gilhooly and Logie (1980a) divided participants into two subgroups with approximately the same number of males and females in each and found a correlation of .98 between them.

The reliability of subjective AoA measurements has also been shown by correlating ratings obtained in different studies. The ratings collected by Gilhooly and Logie (1980a) and Gilhooly and Hay (1977) had 54 word in common; the correlation between the two measurements for those words was .97. Bird, Franklin and Howard (2001) also report a very high correlation between their ratings and Gilhooly and Logie for 81 common words (r = .91).

Simple correlation is not the only technique that has been used in assessing the reliability of AoA estimates. Rubin (1980) found a reliability coefficient of .99 using Cronbach's (1951) alpha, while Carroll and White (1973a, 1973b) reported reliability coefficients of .97 (1973a) and .98 (1973b) using Ebel's (1951) intraclass correlation method.

Authors have gone to great lengths to show the validity of AoA ratings as a representation of the real age at which words are learned. Indeed, many of the objective estimates of AoA described in the previous section were created in order to test the validity of the subjective estimates that are more often used. Carroll and White validated their use by showing a strong correlation (r = .85) between rated AoA and their objective AoA estimates based on age-tagged frequency estimates (see above for more details about these objective estimates).

Gilhooly and Gilhooly (1979) took a two-pronged approach in testing the validity of AoA ratings. In their first experiment, they created a list of words with known objective AoA estimates (derived from the Mill Hill Vocabulary Scale) and compared them with subjective estimates provided by a group of naïve adult raters. The authors found a significant correlation of .93 between the words' rank position in the Mill Hill Vocabulary Scale and the rated AoA estimates. In their second experiment, Gilhooly and Gilhooly tested groups of children of different ages on their knowledge of a list of words with a wide range of subjective AoA ratings taken from the Gilhooly and Hay (1977) study. The objective age of acquisition for each of these words was taken as the average age, in months, at which 50% of participants would know the word. The correlation between the objective AoA estimates thus obtained and the Gilhooly and Hay's subjective estimates was .84. These results lead Gilhooly and Gilhooly to conclude that adult subjective estimates are a valid measure of AoA. Morrison, Chappell and Ellis (1997) also compared AoA subjective ratings of pictureable nouns with the age at which at least 75% of a sample of school-children were able name the corresponding pictures. The correlation between these two estimates of AoA was .75.

De Moor, Ghyselinck and Brysbaert (2000) endeavored to validate the subjective AoA norms they collected in a previous study (Ghyselinck et al., 2000). The authors showed that 6-year olds knew most words rated as early acquired, and that many 12-year olds did not know words rated as late acquired. Furthermore, they showed that words with early AoA ratings have a higher chance of being used in spontaneous speech by 4-year olds.

Jorm (1991) performed a longitudinal study on his own daughter to asses the validity of AoA ratings. He recorded the age at which the child first said and read each of 94 pictureable nouns adapted from Carroll and White (1973). When the child was 9.5 and 11.5 years old, he asked her to estimate, in years, how old was she when she first said each of the words. These estimates showed a very high correlation with both the actual age in which she first said and read the words, with slightly higher correlations for her estimates given at age 11.5. Furthermore, the AoA estimates were very stable across the two different ages at which they were obtained, speaking to the reliability of the ratings over time for the same subject. There was also a high correlation between these estimates and the ones obtained by Carroll and White from adult raters. A regression study showed that AoA ratings were also somewhat influenced by word length (for the ratings at age 9.5) and written word frequency (for the ratings at age 11.5), but the ages at which a word is first said and read were the best predictors of the ratings.

Walley and Metsala (1992) also showed that even AoA estimates given by children are valid approximations to real AoA. In their study, the authors collected ratings from two groups of children (one with 5 year olds and one with 7 year olds) by asking them when they had learned a word, or when they thought they were going to learn them. They compared these ratings with adult estimates of AoA collected for the same sample of words and found that their means were very similar, and that there was a high correlation between the ratings provided by all three age groups. Furthermore, they showed in a regression analysis that both the children's and the adult's ratings were the best predictors for children's performance on tasks that included picture recognition, detection of mispronunciations, and vocabulary monitoring.

In summary, there is a large body of evidence that indicates that subjective estimates of AoA are a reliable and valid method to represent the order in which words are learned.

1.4. Disentangling AoA and Frequency Effects.

A major problem with attempts to isolate the real contribution of AoA to visual word recognition performance is the high correlation between AoA and other lexical variables, and particularly with word frequency; for the most part, words that are learned earlier are also used more frequently throughout life (Brown & Watson, 1987; Gilhooly & Logie, 1981;

Morrison & Ellis, 1995; Rubin, 1980); for example, Carroll and White (1973) reported a correlation of -.68 between these two variables.

Despite the large number of studies that have attempted to determine the relative importance of word frequency and AoA, a definitive answer has proven elusive, with some evidence supporting the frequency account and other reports supporting the existence of independent AoA effects. Methodologically, the two main approaches that have been taken to solve this problem are multiple regression and factorial designs. In both cases, the tight relationship between the characteristics of words has hindered the design of experiments that could settle the issue in a definitive way. As argued by Morrison (2003), results from multiple regression studies should be interpreted with caution because the high correlation between AoA and Frequency may lead one to misattribute variance to the variables involved. In the case of factorial designs, the high correlation between AoA and Frequency narrows the range of words that can be used in each of the conditions, limiting the possibility of controlling for the two variables of interest and other word characteristics that may influence the outcome (Morrison & Ellis, 1995). This is particularly true for Late Acquired/High Frequency words.

As the debate on this issue currently stands, most opinions can be grouped around three positions. A number of studies (e.g., Brown & Watson, 1987; Carroll & White, 1973; Gilhooly, 1984; Gilhooly & Gilhooly, 1979; Morrison & Ellis, 1995; Morrison et al., 1992; Vitkovitch & Tyrrell, 1995) maintain that effects that were previously attributed to frequency can be reduced to Age of Acquisition effects, at least for certain word recognition tasks. Another position defends that, on the contrary, frequency is the most important determinant in lexical access and that AoA effects can be explained in terms of their "cumulative frequency" (Lewis, 1999a, 1999b; Zevin & Seidenberg, 2002, 2004). Lastly, some studies support independent effects for both AoA and frequency (e.g., Barry et al., 1997; Brysbaert et al., 2000; Ellis & Morrison, 1998; Gerhand & Barry, 1998, 1999a, 1999b). More recently, the debate about the relative importance of AoA and frequency has been re-fuelled by a paper by Zevin and Seidenberg (2002) that pointed out methodological flaws in several previous studies that reported AoA effects. Therefore, at present, it is unclear whether AoA contributes independently from frequency.

Determining the relative influence of AoA and Frequency and the way they interact with each other is of paramount importance in the elaboration of theoretical and simulation models of word recognition. From an empirical standpoint, it is important to determine which of the two variables should be controlled for in studies involving word recognition (Gerhand & Barry, 1998). For theoretical purposes, it is vital to determine whether the numerous models of word recognition that rely heavily on frequency accounts are correct or if they should be modified in favor of AoA-compatible models.

1.5. The Focus of this Thesis.

The main purpose of the present thesis is to provide a better characterization of Age of Acquisition and frequency effects in visual word recognition. As has been shown in the introduction, these two effects have come to dominate much of the current debate in the visual word recognition literature, and they both provide strong constraints for models that attempt to explain how words are read. Answers about whether either of these effects are truly present and how they are present could help sort out models that are adequate from the ones that are not.

The rest of this dissertation is structured as follows:

Chapter 2 presents the *Bristol Norms for Age of Acquisition, Imageability and Familiarity*, a large set of ratings for more than 1,500 words. One of the stumbling blocks in the research of AoA and frequency effects has been the lack of databases large enough to allow the selection of adequate stimuli sets while controlling for other relevant variables. These ratings will then be used to replicate seminal findings concerning AoA and word frequency effects while addressing some methodological concerns that have been mentioned in the literature. The norms will also be used to explore whether familiarity measures provide a valid representation of word frequency.

Chapter 3 presents a series of experiments that explore whether AoA and frequency have independent effects in visual word recognition, and particularly, whether AoA effects can be reduced to cumulative frequency effects or not. This is done using "expert vocabularies", that is, words that were learned quite late in life but have very high frequency for a specific set of people (i.e., psychologists, chemists or geologists).

Chapter 4 turns to frequency and considers whether spoken word frequency has an effect on visual word recognition or not. This issue has received surprisingly little attention in the literature, and could have a major impact in evaluating models of reading. Words with "unbalanced" frequencies (i.e., high spoken frequency but low written frequency or vice versa) were compared with words with similar frequencies in both modalities

In combination these studies provide a collection of constraints that can be used to evaluate, and indeed shape, models of visual word recognition.

Chapter 2. The Bristol Norms for Age of Acquisition, Imageability and Familiarity

Chapter 2

THE BRISTOL NORMS FOR AGE OF ACQUISITION, IMAGEABILITY AND FAMILIARITY[1].

2.1. General Background

Previous research has identified a number of variables that affect the speed and accuracy with which words can be recognised, recalled, named, and/or classified (e.g., Balota et al., 2004; Hulme, Stuart, Brown, & Morin, 2003; Roodenrys, Hulme, Alban, Ellis, & Brown, 1994; Whaley, 1978). Some of these variables, such as word length and word onset, are intrinsic to each word and can be determined directly from their surface structure without reference to any other materials. Other variables, such as neighbourhood size, word frequency, bigram frequency, or regularity depend on the relationship of the target item to a larger corpus of words; these values are estimated by placing the word within a certain linguistic context. Obtaining data for these two types of variables is, in general, not problematic: given the existence of a reasonably comprehensive word frequency corpus, such variables can be obtained for virtually all of the words in a language, and various software tools exist to facilitate the retrieval of these statistics (e.g., Davis, 2005; Davis & Perea, in press; Duyck, Desmet, Verbeke, & Brysbaert, 2004). However, there are also other variables that affect the speed and accuracy of word recognition and recall that are a reflection of the personal experiences of language users. Such variables include Age of Acquisition (AoA), imageability, and familiarity. Measurements for this type of variable are usually estimated by asking people to make subjective ratings (e.g., "How old do you estimate you were when you learned this word?").

[1] Adapted from Stadthagen-Gonzalez and Davis (in press)

Collecting norms for these subjective variables is a relatively time-consuming process, and most norming studies have been limited to the number of items that a participant can rate in one session of reasonable length. Most of the published norming studies concentrate on only one or two of the three variables that are the focus of the this chapter, and typically include less than 1000 items.

The limited availability of ratings for variables like AoA, imageability, and/or familiarity poses a particular problem to researchers. When designing experimental stimuli, experimenters who wish to match stimuli on subjective variables like AoA must typically run their own norming study on candidate stimuli. More frequently, experimenters simply forgo the possibility of matching stimuli on these subjective variables (and hope for the best). The limited availability of norms for subjective variables also hampers the research strategy of performing multiple regressions on large-scale databases. For example, Balota et al.'s (2004) recent regression analyses of speeded naming and lexical decision latencies for 2,428 monosyllabic English words omitted AoA as a predictor variable because norms were only available for a quarter of the items in their set.

The present chapter aims to improve this situation by increasing the number of words for which subjective norms are available. The *Bristol Norms* presented here consist of ratings of AoA, imageability, and familiarity for 1,526 words. These particular variables were chosen for this study because of the interest that each of them has generated in the language and memory literature in recent years, either as the explicit focus of study or as extraneous variables to control.

Age of Acquisition refers to the age at which a word was learnt and has been proposed as a significant contributor to language and memory processes (e.g., Carroll & White, 1973; Hirsh & Funnell, 1995; Juhasz & Rayner, 2003; Morrison et al., 1992; Roodenrys et al., 1994).

Although some studies have used "objective" AoA norms (Morrison et al., 1997), most experimenters choose to use subjective measures of AOA. These AoA estimates have been shown to be reliable (Carroll & White, 1973; Gilhooly & Logie, 1980a) and to provide a valid estimate for the objective age at which a word was acquired (De Moor et al., 2000; Gilhooly & Gilhooly, 1980; Morrison et al., 1997).

Imageability is a semantic variable that measures how easy it is for a word to arouse mental images. It has been used to evaluate the effect of meaning on memory and word recognition (Balota, 1990; Balota et al., 2004; Paivio, 1971; Paivio, Yuille, & Madigan, 1968). Imageability is closely related to concreteness: for most words the two measures are quite similar, although there are some exceptions. Bird, Franklin, and Howard (2001) give the example of the word *armadillo*, which generates high concreteness but low imageability ratings, presumably because of a lack of personal exposure to armadillos. Some studies suggest that imageability is a better predictor of performance than concreteness (Bruyer & Strypstein, 1985; Marcel & Patterson, 1978; Richardson, 1975).

Familiarity ratings have often been interpreted as a measure of the frequency of exposure to a word, and Gernsbacher (1984) suggested that familiarity is a better predictor of word performance than printed word frequency, particularly for low-frequency words. However, it is not entirely clear what processes are involved when readers make ratings of the familiarity of a word. For example, it has also been argued that familiarity ratings include a semantic component (Balota, Pilotti, & Cortese, 2001). The present approach of obtaining ratings of both subjective familiarity and imageability for a large sample of words (for which printed word frequency estimates are also available) allows further investigation into this issue.

The relevance of each of these three variables has been challenged at one point or another. On the one hand, Zevin and Seidenberg (2002) have argued that effects that are purportedly due to AoA actually reflect the cumulative frequency of exposure to a word (though see Chapter 3 of the present dissertation, for recent experimental evidence undermining this cumulative frequency hypothesis). On the other hand, after controlling for AoA, Morrison and Ellis (2000) failed to find effects of either imageability or familiarity on word naming latency. Finally, although subjective familiarity has been used widely, its usefulness has been strongly challenged: Brown and Watson (1987) questioned Gernsbacher's (1984) claim that familiarity is a more suitable frequency measure than objective frequency, and Balota et al. (2001) proposed that subjective frequency is a more useful estimate than subjective familiarity. These three variables were chosen precisely because there are several lines of debate surrounding them, and this generates a need for normed stimuli that can be used to explore each position. Indeed, the norms presented here would be equally useful for studies trying to prove as well as disprove the relative importance of each of these variables.

To date, the largest database of ratings for AoA, imageability, and familiarity in English is the set compiled by Gilhooly and Logie (Gilhooly & Logie, 1980a; henceforth G&L), which includes norms for these variables for 1,944 words, as well as measurements for concreteness and ambiguity. This database has proved to be a valuable resource for experimenters: the G&L norms have been referenced by at least 200 papers since their publication. The G&L norms are also incorporated into the MRC Psycholinguistic Database (Coltheart, 1981). In order to provide an even wider choice of words available to experimenters, the Bristol Norms were designed to be compatible with the G&L norms with respect to the definition of the variables as well as the composition of the sample of participants (i.e., mostly young undergraduate students). Analyses reported later on show that it is legitimate to merge the two databases. The Combined Bristol/G&L Norms provide normative data for 3,394 words, the largest set of norms for those variables so far in English. A wider availability of items will be of great help in the implementation of factorial designs or regression studies in the fields of language, memory, and neuropsychology. This chapter also proposes a modular approach to collecting norms, which enables incremental increases in the size of the database.

2.2. The Bristol Norms

2.2.1. General Method

Participants

Ratings from 100 participants were obtained for each of the variables under study (AoA, Imageability, and Familiarity). All participants were undergraduate students enrolled in level 1 and 2 psychology courses at Bristol University. They were all native British English speakers and their average age was 19.7, with a range of 18 to 40 years.

The Word Corpus

A set of 1,526 words was rated in this study. The selection of words was governed by two main criteria, both of which aimed to maximise the usefulness of the norms. The first of these was that the words should be fairly representative of the types of words typically used in psycholinguistic experiments. Consequently, the Bristol Norms include words that are relatively short (between 4 and 7 letters, with one or two syllables), and excluded regular past tenses and plurals; some irregular past tense and plural forms were included (e.g., *bought*, *teeth*). Furthermore, although the full sample covered a range of frequencies from 0.34 to 1,642 counts per million in the CELEX database (Baayen et al., 1995), the focus was placed on words in the frequency range most often sampled by psycholinguistic experiments; thus, 93% of the words had CELEX written frequencies of between 1 and 100 counts per million. The second criterion for inclusion in the set of words to be rated was that the word was not already included in the G&L norms. This resulted in a set of 1,450 words. In addition, a set of 76 words that also had AoA, imageability, and familiarity ratings in the Gilhooly and Logie (1980a) norms were included as reliability controls.

Procedure

The 1,450 new words were randomly divided into 5 sets of equal size and the 76 control words were added to each set. Each block of 366 words was presented in four columns on a computer spreadsheet and rated on one of the three variables (AoA, imageability, or familiarity) by 20 volunteers in sessions that lasted no more than 30 minutes. The words were presented in a different random order for each group of 4 participants. This procedure was repeated for AoA, imageability, and familiarity, with the pertinent changes in the instructions for each variable as outlined below. The exact wording of the instructions presented to raters for each variable are presented in Appendix A.

Raters for *AoA* were asked to type next to each word an estimate in years of when they learned the word. It was decided that asking participants to provide their ratings in years was simpler than using the more complex scale method favoured by other rating studies; Ghyselinck, De Moor, and Brysbaert (2000) have shown that there is no real difference in results if participants are asked to rate words by using a 7-point scale or by entering their estimates of AoA in years. In order to make the Bristol Norms compatible with the G&L Norms, responses were then converted to the same 1 to 7 scale used there (from 1 for age 0 to 2 years, to 7 for age 13 years or more, with intervening bands spanning 2 years).

Raters for *imageability* were asked to indicate how easily each word elicited mental images. They indicated their answer on a scale from 1 to 7, where 1 corresponded to low imageability and 7 indicated high imageability. The instructions given to participants were almost identical to those used by G&L, which were based on the instructions devised by Paivio et al. (1968). The only differences in wording related to the method by which raters made their responses. Instead of circling the chosen number on the scale, participants typed it on a space given next to each word; the 7-point scale was visible at all times at the top of the screen.

Raters for *familiarity* were also asked to provide ratings using a 7-point scale, with 1 being assigned to words that they never had seen and 7 to words that they had seen very often (nearly every day). The wording of the instructions for this task was also very similar to the ones provided by G&L. As in the case of the imageability ratings, participants typed the appropriate number next to each word, with the 7 point scale visible at all times.

In all cases, ratings on the 1-7 scale were subsequently multiplied by 100 and rounded to the nearest integer so as to be able to present all the ratings as integers on a scale from 100 to 700.

Results and Discussion

The first step in the analysis involved combining the five separate blocks into a single data set. The 76 control words (which were rated by all participants) were used to verify the validity of this approach and to homogenize the ratings across blocks. The ratings for control words had interblock correlations of at least .91 across all three variables, and for all three variables the inter-rater reliability coefficient (Cronbach's alpha, 1951) was at least .98. The linear transformation procedure outlined by Coltheart (1981; Appendix 2 or the MRC Database Handbook) was followed in order to homogenize the means and standard deviation across the five blocks (the same procedure was used by Bird et al., 2001 to transform their norms to be on the same scale as the G&L norms). This involved three steps. First, the overall mean rating was computed (across all 100 participants) for each of the 76 control words. These overall means were then used to predict the mean rating (across 20 participants) for each of the five blocks, resulting in five separate regression equations. For each block of 366 words the raw mean rating of each item was then transformed by subtracting the intercept of the regression equation for that block and dividing by the regression coefficient. The five blocks were then combined into a single data set; for the control words, the value entered into the database was the average of the transformed ratings for all five blocks.

The reliability of the resulting norms was assessed by examining the correlations between the ratings for the 76 control words in the Bristol Norms and the corresponding ratings for the same words in the G&L Norms. These correlations were very high (AoA: $r = .89$; IMG: $r = .93$; FAM: $r = .86$), and the reliability index between the Bristol Norms and the G&L norms was also very high (AoA: alpha = .93; IMG: alpha = .96; FAM: alpha = .86). There was a very high correlation between the imageability norms presented here and those recently reported by Cortese and Fugett (2004) ($r = .84$, $N = 680$). The correlations with the standardized reaction times for lexical decision and word naming taken from the English Lexicon Project (ELP; Balota et al., 2002) were also examined for the 76 control words. For lexical decision latencies, the correlations were as follows: AoA $r = -.27$, IMG $r = -.13$, and FAM $r = -.39$; for naming latencies the correlations were AoA $r = -.08$, IMG $r = -.16$, FAM $r = -.10$. These correlations were very similar to those between the latency variables and the G&L norms across the same set of words (e.g., the correlation between G&L AoA and lexical decision latency was -.26). There were no significant differences between the correlations for the two sets of norms (all p values >.24 using the procedure outlined by Meng, Rosenthal, & Rubin, 1992).

Across the entire sample, the correlations between each of the three subjective variables and standardized reaction times for lexical decision and word naming taken from the ELP (Balota et al., 2002) were highly significant (see Table 2.1).[2] Note that the correlations with the reaction time measures were particularly strong in the case of the AoA and familiarity ratings. This offers some support for the validity of the collected ratings.

[2] Twenty-three items included in the Bristol Norms did not have entries in the Balota et al. (2002) database, mainly because of differences in spelling between the British and the American dialects (e.g., *favour* vs. *favor*).

Table 2.1: *Correlations Between Subjective Ratings From the Bristol Norms and Reaction Time Measures.*

	Lexical Decision	Naming
AoA	0.51*	0.33*
IMG	-0.22*	-0.13*
FAM	-0.53*	-0.33*

Note: Lexical Decision and Naming reaction times correspond to the standardized latencies reported in the ELP database (Balota et al., 2002).; N = 1503
AoA = Age of Acquisition; IMG = Imageability; FAM = Familiarity
$p < .001$

2.2.2. Correlations Between the Subjective Norms and other Lexical Variables.

Table 2.2 shows the correlations between AoA, imageability, and familiarity and a selection of lexical variables: word length (in letters, phonemes, and syllables), mean logarithmic bigram frequency (MLBF), neighbourhood size (N), and a variety of measures of written and spoken frequency: CELEX written frequency (Baayen et al., 1995), British National Corpus (BNC) written frequency (Aston & Burnard, 1998), BNC spoken frequency (based on the *BNC Demographic Database*), and the *Educator's Word Frequency Guide* (WFG; Zeno et al., 1995).

As can be seen, AoA is significantly correlated with each of these variables. The direction of these correlations accords with expectations regarding the age at which words are acquired. Thus, words that are acquired later tend to be longer, less frequent, less imageable, less subjectively familiar, and have fewer neighbors and lower bigram frequencies than words that are acquired earlier. To investigate the independent contribution of each of these variables, a simultaneous multiple regression was conducted with AoA as the dependent variable and six independent variables (see Table 2.3). To avoid problems of excessive multicolinearity among the independent variables, only a single measure of written word frequency (the WFG count) and a single measure of length (number of phonemes) were used. As can be seen in

Table 2.2: *Correlations Between Subjective Ratings From the Bristol Norms and Other Lexical Variables (Length, Bigram Frequency, N, and Frequency).*

	Variable	AOA	IMG	FAM
1	AOA	+1.00	-	-
2	IMG	-0.53	+1.00	-
3	FAM	-0.61	+0.12	+1.00
4	LEN_L	+0.30	-0.21	-0.13
5	LEN_S	+0.35	-0.22	-0.15
6	LEN_P	+0.34	-0.22	-0.15
7	MLBF	-0.21	+0.05 ns	+0.14
8	N	-0.29	+0.18	+0.11
9	log10 (CELEX Written+1)	-0.48	+0.02 ns	+0.60
10	log10 (BNC Written+1)	-0.38	-0.05 ns	+0.57
11	log10 (WFG+1)	-0.54	+0.09	+0.57
12	log10 (BNC Spoken+1)	-0.66	+0.25	+0.72

Note: *ns* = not significant at the .05 level. All other correlations are significant at the p <.001 level. AoA = Age of Acquisition; IMG = Imageability; FAM = Familiarity; LEN_L = length in letters; LEN_S = length in syllables; LEN_P = length in phonemes; MLBF = mean log bigram frequency by tokens ; N = orthographic neighbourhood.

Table 2.3, all six variables included in the regression made independent contributions to predicting rated AoA, with the best predictors being imageability and familiarity, followed by spoken and written frequency. This agrees with the conclusions drawn in previous investigations of AoA based on different sets of words (e.g., Bird et al., 2001; Gilhooly & Logie, 1980a; Morrison et al., 1997).

Imageability is significantly correlated with AoA, familiarity, length, and N, but less clearly with word frequency. Thus, more imageable words tend to be acquired earlier, are more familiar, tend to be shorter, and tend to have more orthographic neighbours than less imageable words. It is unlikely that the correlation with N has any causal component; rather, it probably reflects the fact that shorter, earlier-acquired words have more neighbours than longer, later-acquired words. When AoA and length are partialled out, the partial correlation

Table 2.3: *Multiple Regression Analysis With Rated Age of Acquisition as the Dependent Variable and Six Independent Variables.*

Variable	Coefficient			
	B	*SE*	β	t
Imageability	-.373	.015	-.409	25.17
Familiarity	-.675	.044	-.342	15.17
Spoken Frequency	-44.672	5.491	-.210	8.14
Written Frequency	-22.897	5.654	-.087	4.05
Number of Phonemes	8.531	2.084	.077	4.09
N	-1.915	.569	-.062	-3.37

Note: Spoken Frequency = log 10 (BNC Spoken frequency + 1),
Written Frequency = log 10 (WFG frequency + 1).
All t values are significant at $p < .005$.

between N and imageability is -.01 (i.e., non-significant, negligible, and in the opposite direction to the raw correlation). With respect to the correlations between imageability and frequency, one of the three measures of written frequency (the CELEX written frequency count) shows a negligible correlation with imageability, another (the BNC written frequency count) shows a trend toward a negative correlation with imageability (p = .051), and the third (the WFG) shows a significant positive correlation with imageability; the latter correlation may reflect AoA, given that the WFG count is based exclusively on a corpora of school texts. It may be concluded that any correlation between imageability and written frequency is, at best, very weak. Likewise, the positive correlation between imageability and spoken frequency probably depends on the strong correlations with the third variable of AoA. When AoA is partialled out, the partial correlation between imageability and spoken frequency is negative (-.15, i.e., high imageability words tend to be lower in spoken frequency).

Some authors (Zevin and Seidenberg, 2002, 2004) have proposed that AoA effects in visual word recognition actually reflect the effect of cumulative frequency, i.e., words that are acquired earlier in life will have been encountered more often overall, when equating for printed word frequency. Zevin and Seidenberg (2004, 2002) calculated cumulative frequency as the sum of frequency estimates for all grades included in the Zeno et al (1995) norms. We computed cumulative frequency in the same way for the 1307 words in our sample that are listed in the Zeno et al. database. The correlation between cumulative frequency and rated AoA was relatively high (r = -0.24), which is at least consistent with the possibility that previously observed effects that have been attributed to AoA could have been the result of a confound with cumulative frequency. However, the correlations between cumulative frequency and lexical decision (r = -0.21) and word naming latencies (r = -0.14) were much weaker than for the AoA subjective estimates included in the Bristol Norms: Rated AoA explains about 25% of the variance for lexical decision latencies, whereas cumulative frequency explains only about 4% of this variance. This difference leads us to conclude that the effects of rated AoA on visual word recognition performance are not simply due to cumulative frequency. Experimental investigations of the cumulative frequency hypothesis have reached the same conclusion (Ghyselinck, Lewis, & Brysbaert, 2004; Stadthagen-Gonzalez, Bowers, & Damian, 2004).

2.2.3. What Does Subjective Familiarity Measure?

One question that the norms may help to address is exactly what is being measured by subjective familiarity ratings. As noted already, it has been suggested that these ratings may provide a better measure of the relative frequency of exposure to a word than objective measures of printed word frequency (Gernsbacher, 1984; Gilhooly & Logie, 1980a). The Bristol norms show relatively strong correlations between familiarity and both written and spoken frequency (see Table2.2), which supports the idea that subjective familiarity ratings reflect frequency of exposure. One way to assess familiarity ratings provide a better measure of the relative frequency of exposure to a word than objective measures of printed word frequency is to examine how the correlations between familiarity and objective frequency measures compare with the intercorrelations between different objective frequency measures. Theoretically, variance in measures of word frequency can be partitioned into two separate components: a systematic component, reflecting "true" word frequency, and a random component, reflecting measurement error. The measurement error component varies across metrics, with some frequency measures containing greater error variance than others. The correlation between different frequency measures will decrease as a function of the magnitude of the error variance in the two measures (e.g., there will be a correlation of 1 for two measures with zero error variance, and a correlation approaching zero if one or both of the measures has extremely large error variance). Thus, if subjective familiarity offers a better (e.g., less noisy) measure of frequency of exposure than objective measures it should correlate more highly with these objective measures than the objective measures correlate with each other. Thus, if subjective familiarity offers a better (e.g., less noisy) measure of frequency of exposure than objective measures it should correlate more highly with these objective measures than the objective measures correlate with each other. This is clearly not the case for objective measures of written frequency.[3] As Table 2.2 shows, the maximum correlation between subjective familiarity and any of the objective written frequency measures was .60. By contrast, the minimum correlation between the objective written frequency measures was .83; in the case of the two objective measures that were based on British English (the CELEX and BNC counts) the correlation was .91. This agrees with findings derived from entirely distinct sets of words (Brown & Watson, 1987; Gordon, 1985), and implies that objective printed frequency measures are more valid measures of the frequency with which readers encounter a word in print. Furthermore, the correlation between the Bristol

[3] However, the correlations between objective and subjective measures of word frequency could be slightly diluted due to the difference in ranges for these two types of measures.

subjective familiarity norms and those from the MRC database (of .65) is greater than any of the correlations with objective written frequency measures.

There is some support for the possibility that subjective familiarity is a measure of the spoken frequency of the word. The correlation between the Bristol familiarity norms and the BNC spoken frequency measure is .72, which is slightly greater than the correlation of .67 between the BNC and CELEX (objective) spoken frequency counts. However, the correlation between the Bristol familiarity norms and the CELEX spoken frequency measure is only .58; the weaker correlation in this case probably reflects the poorer reliability of the CELEX spoken frequency count, which is based on a much smaller sample of speech than the BNC count. The possibility that familiarity is tapping into spoken rather than written frequency is also supported by multiple regression analyses in which familiarity is the criterion variable. Log BNC written frequency explains 33% of the variance in the familiarity norms, but adding log BNC spoken frequency to the equation accounts for an additional 20% of variance (R^2 = .53). By contrast, when the order of entry of these variables is reversed, log BNC spoken frequency accounts for 51% of variance in familiarity by itself, and the addition of log BNC written frequency explains only an extra 2% of variance. It should be noted that (following Gilhooly & Logie, 1980a) the instructions for rating familiarity specified that raters should take into account both the frequency with which they had seen and heard the word in question, and that they should rate the word on the basis of the higher of the two measures. The majority of individuals probably hear many more words than they read, and so it is not surprising that subjective estimates of word familiarity should be biased towards frequency in the spoken rather than written modality.

A related question concerns whether subjective familiarity ratings tap into anything beyond frequency information. As Balota (2001)noted, familiarity ratings may also reflect variables unrelated to word frequency, such as the meaningfulness of the word (c.f., Toglia & Battig, 1978) or the familiarity of the sublexical spelling-sound correspondence. Brown and Watson (1987) noted that subjective familiarity was strongly correlated with AoA. The correlations in Table 2.2 show that the Bristol familiarity norms correlate most strongly with spoken frequency, written frequency, and AoA, and relatively weakly with imageability. Together, the two BNC log frequency measures and the Bristol AoA norms account for 57.3% of the variance in familiarity ratings (N = 1526). The addition of imageability increases this to 59.3%. However, the partial correlation with imageability is -.22; that is, once frequency and AoA have been partialled out, there is a negative correlation between imageability and familiarity, which is opposite to the direction that might be expected.

A final question regarding familiarity concerns how useful it is as a predictor of the speed of word identification. By itself, familiarity accounts for 28% of the variance in standardized lexical decision latencies from the ELP database and 10% of the variance in standardized naming latencies. This is not especially surprising, given the very high correlation of familiarity with written frequency, spoken frequency, and AoA. Once these three variables are partialled out (using the BNC frequency counts) the residual variance in subjective familiarity explains only 1.4% of the variance in standardized lexical decision latencies from the ELP database and only 0.4% of the variance in standardized naming latencies. In summary, then, subjective familiarity appears to be inferior to objective frequency counts as a measure of the frequency with which words are encountered in print, although it may offer a reasonably good measure of the frequency with which words are encountered in speech. To the extent that subjective familiarity taps into something beyond frequency, it appears to offer no advantage over more clearly defined measures such as objective frequency or AoA. On this basis, it seems pertinent to question the usefulness of this variable in psycholinguistic research. The previous analysis coincides with the findings presented by Baayen (Baayen, in press) who used a variety of statistical analyses to characterize the influence of a several predictors (such as word frequency, length and neighborhood) over visual lexical decision latencies, word naming latencies and subjective familiarity ratings. The results from these analyses lead Baayen to conclude that "the observation that familiarity ratings are an independent variable in their own right, just as response latencies or eye fixation durations, has important methodological consequences. Ratings should not be used as a substitute for corpus-based frequency counts. Matching for familiarity ratings, for instance, implies at least partial matching for a series of other variables, potentially including variables of interest, and reduces the likelihood of finding significant effects. Likewise, familiarity ratings should not be included along with frequency counts in a regression analysis of, for instance, lexical decision latencies [...] just as one would not normally include lexical decision latencies as a predictor for, e.g., eye fixation durations" (p. 6).

Nevertheless, subjective familiarity ratings may be appropriate for use in research in other areas (such as memory experiments and neuropsychological case studies), when experimenters are seeking to control their stimuli on a very limited set of variables.

2.2.4. Merging the Bristol Norms with the Gihooly & Logie (1980) Norms

In view of the evidence that the Bristol and G&L databases are quite compatible, and given the fact that the instructions used were very similar to those used by G&L, it was decided to form a "megadatabase" that merged the Bristol norms with the G&L norms, resulting in a set of norms for a total of 3,394 words. Other large norming studies were also considered as candidates to be merged into this megadatabase, but only G&L was deemed compatible enough with the Bristol Norms for a variety of reasons. For example, the control words in the Bristol Norms were highly correlated with the norms for AoA and imageability collected by Bird, Franklin and Howard (2001). However, the correlation between the Bird et al. ratings and the ELP latencies for lexical decision and naming were relatively poor (for Lexical Decision: AoA, $r = .12$; IMG, $r = .00$; for Naming: AoA, $r = .03$; IMG, $r = .06$). This result may be attributable to two characteristics of the Bird et al. norms. First, raters in that study were considerably older than the ones that participated in the Bristol Norms and G&L, as well as the participants in the lexical decision and naming tasks reported by Balota et al. (2002). Secondly, words in the Bird et al. norming study were unambiguously presented as verbs or nouns, a disambiguation that is not possible in single-word recognition experiments.

The Bristol norms and the G&L norms were merged after applying the linear transformation procedure already described above (i.e., transforming the Bristol norms based on the coefficients of a regression equation predicting these norms on the basis of the G&L norms). Some ratings that were already near the extreme values of the scales (100 and 700) were pushed outside the scale by the linear transformation, so the values for those items were set to the extreme values of 100 or 700 according to each case. There were 27 such items for AoA, 6 for imageability and 1 for familiarity. The mean transformed ratings for AoA (both in years and in the 100-700 scale), imageability, and familiarity for all of the Bristol norms can be downloaded from http://language.psy.bris.ac.uk/bristol_norms.html.

A methodologically interesting aspect of this study that is worth noting is the use of a control set of words to enable a modular approach to the collection of subjective norms. The high reliability and validity of the Bristol norms demonstrates the feasibility of assembling large norms from modules of words, provided that both the instructions for each rating and the pool of participants are similar, and that an adequate quality control process is applied by assessing the inter-rater reliability of each block for a set of common words as if each block was a judge evaluating the same items. Future application of this methodology will allow further increases in the size of the database, with the potential goal of providing subjective norms for the entire set of words that are likely to be used in psycholinguistic or memory experiments.

In summary, the present study provides a very large set of norms for variables that are currently relevant to diverse lines of research in the fields of language and memory. It is expected that this large norming study, integrated with the G&L norms, will prove a valuable resource in facilitating experimental research in those fields.

2.3. Application of the Bristol Norms.

Many earlier studies that looked into AoA and frequency effects used the Kucera and Francis (1967) norms as the basis for their estimations of word frequency. These norms are based on the Brown corpus, a sampling of edited English texts printed in the United States during 1961 with a little over one million words. Several authors have pointed out that this relatively small sample size compromises the accuracy of frequency counts, particularly for low frequency words. For example, Balota, Cortese, Sergent-Marshall, Spieler and Yap (2004) conducted a comparison of different word-frequency estimates, and found that the Kucera and Francis accounted for the least amount of variance for word naming and lexical decision latencies. These concerns about the accuracy of the Kucera and Francis norms lead Zevin and Seidenberg (2002) to call into question many influential studies that reported AoA effects in lexical decision (Gerhand & Barry, 1999b; Morrison & Ellis, 1995; Turner et al., 1998) and word naming (Gerhand & Barry, 1999b; Lund & Burgess, 1996; Monaghan & Ellis, 2002; Morrison & Ellis, 1995). When Zevin and Seidenberg re-estimated frequencies using larger databases such as CELEX and the Educator's Word Frequency Guide, they found that words in the early-acquired conditions were often higher in frequency than late-acquired items, suggesting that their advantage was the product of frequency differences rather than

AoA. They also found that, in those studies, early and late items also differed significantly on familiarity, which some authors claim is a better estimate of word frequency, particularly for low frequency words (Gernsbacher, 1984). This could mean that many of the seminal experiments that showed AoA effects could be tainted by this confound with frequency. These methodological problems on past studies lead Zevin and Seidenberg (2002) to conclude that "the evidence for an effect of AoA on skilled reading is weak at best" (p. 2).

As a first step in furthering the exploration of the influence of AoA on visual word recognition, it would seem wise to replicate the basic findings of these earlier studies, while removing the possible frequency confound pointed out by Zevin and Seidenberg (2002) by including several frequency measures as well as familiarity ratings. Great effort was made to tightly match conditions on a variety of relevant variables.

The experiments presented below adopted a semi-factorial design similar to the one used by Morrison and Ellis (1995). In that study, the authors manipulated AoA while frequency was held constant and vice versa. Morrison and Ellis found a significant AoA effect when word frequency and other relevant variables were matched between conditions, but no frequency effect when AoA was controlled for. The AoA effect found in this way was later reproduced by Gerhand and Barry (1998), but they found a significant frequency effect as well.

Experiment 1 examines the contribution of frequency when items are matched on AoA as well as other variables, while Experiment 2 examines the contribution of AoA while frequency is held constant between conditions. The use of the combined Bristol/G&L norms, which provide ratings for a large set of words, allowed me to the assemble a large number of items for each condition, which adds to the generalizeability of the results. About 40% of the items used in both experiments were taken from the newly collected Bristol norms.

Reaction times for word naming and lexical decision for both experiments were obtained from the English Lexicon Project (ELP; Balota et al., 2002), a very large online database that provides behavioural data from visual lexical decision and naming studies, as well as normative data for descriptive variables such as orthographic neighbourhood, length, and word frequency of 40,481 words. The reaction times and error rates in the database were collected in blocks of about 600 words each from groups of students in different universities in the United States. This database is a valuable resource in language research as it provides the opportunity to design ad hoc studies to test hypotheses without spending time and resources collecting empirical data. The use of this approach can be particularly useful for pilot studies, or for experiments such as the present one, in which the main purpose is to confirm past results under better controlled conditions. Balota et al. (2004), Zevin, Seidenberg (2004), and Baayen (Baayen, in press), among others, have used reaction times from this database to conduct a very large regression study on the factors that influence performance on single word recognition. In the present chapter, the data in the ELP database was used in a matched list design.

2.3.1. Experiment 1. Effect of word frequency for words matched on AoA.

Method

Material

There were two conditions with 35 items each: one with high-frequency words ("HiF" condition) and one with low-frequency words ("LoF" condition). Low-frequency items had a frequency count of no more than 15 and high-frequency items had of at least 75 counts per million according to the Educator's Word Frequency Guide. Following Zevin and

Seidenberg's (2002) recommendation, more than one estimate of frequency was used in this experiment, so the two conditions had significant differences in frequency as measured by three different large databases: the Educator's Word Frequency Guide, CELEX, and the British National Corpus. Word frequencies were compressed using the formula *log (1+F)*, where *log* is the base 10 logarithm and *F* is the raw frequency in tokens per million from each of the databases used. Items in both conditions were matched on Age of Acquisition, familiarity and imageability ratings taken from the combined Bristol/G&L norms. They were also matched on word length (measured in number of letters) and Colthart's N measure of orthographic neighbourhood size (Coltheart et al., 1977). Orthographic neighbourhood size refers to the number of words that can be created by changing one letter of the stimulus word while keeping the other letters the same (e.g., *house* and *horde* are orthographic neighbours of *horse*). For the present experiments, the neighbourhood was calculated using the N-Watch software (Davis, 2005). Table 2.4 presents a summary of word characteristics for each condition and *p* values for the t-test between conditions for each of the relevant variables mentioned above. Appendix B shows a complete list of the items included and individual values for each variable.

As can be seen in Table 2.4, whatever trends occurred were not systematically in favour of one group of items over another.

Reaction Times.

The ELP database provides mean reaction times for lexical decision and word naming in milliseconds. It also provides those RTs as z-scores standardized on each participant's overall reaction times. Each data point corresponds to the mean performance across at least 25 participants. A more detailed explanation on how these latencies were obtained can be

Table 2.4: *Summary of item characteristics for Experiment 1.*

	AoA	IMG	FAM	LEN	N	Log WFGF	Log CelexF	Log BNCF
LoF	3.37	5.06	5.54	5.46	2.66	0.83	0.96	0.81
HiF	3.30	4.98	5.64	5 74	3.4	2.2	2.2	2.27
p	.81	.80	.57	.36	.38	< .01	< .01	< .01

Note: AoA = Age of Acquisition; IMG = Imageability; FAM = Familiarity; LEN = length in letters; N = orthographic neighbourhood; WFGF = word frequency according to the Educator's Word Frequency Guide; CelexF = word frequency according to CELEX; BNCF = word frequency according to the British National Corpus.

found in Balota et al. (2004; pp. 286-287). It is important to note that the selection of items for each condition was carried out blind from these reaction time data.

Results.

As mentioned before, reaction times were obtained from the ELP database. Mean reaction times, mean z-transformed reaction times and error rates for Experiment 1 are presented in Table 2.5. Appendix B presents detailed information on an item-by-item basis. Significance tests were performed on the z-transformed reaction times and, given the nature of the ELP data, all analyses presented are by items.

Table 2.5: *Summary of results for Experiment 1.*

	LD_RT	LD_Z	LD_Acc	NMG_RT	NMG_Z
HiF	603	-0.64	0.98	605	-0.55
LoF	659	-0.46	0.97	630	-0.44

Note: LD_RT = mean reaction times for lexical decision; LD_Z = standardized z-scores for lexical decision; LD_Acc = accuracy rate for lexical decision; NMG_RT = mean reaction times for word naming; NMG_Z = standardized z-scores for word naming.

For lexical decision, latencies to more frequent words (e.g. *water*) were significantly lower than for infrequent words (e.g. *sober*) ($F(1,68) = 16.21$; $p < .01$) with no trade-off in accuracy ($F(1,68) = 0.91$; $p = .34$). Results for word naming in this experiment should be interpreted with care, since words in each condition were not matched for phonological onset, a factor that can affect the accuracy of voice triggers (Kessler, Treiman, & Mullennix, 2002; Rastle & Davis, 2002). While holding this limitation in mind, it is still interesting to observe that high frequency words hold a significant advantage in performance over low frequency words ($F(1,68) = 4.02$; $p < .05$), but this difference is much smaller (by nearly half) than the one obtained in the lexical decision task.

2.3.2. Experiment 2. Effects of AoA for words matched on frequency.

Method.

Material

There were two conditions with 35 items each: one with early-acquired ("EarlyAoA" condition) and one with late-acquired words ("LateAoA" condition). Early-acquired words had an AoA rating of no more than 250 and late-acquired words had a rating of at least 500 according to the combined Bristol/G&L norms. Word frequency was closely matched

between the two conditions according to each of the three databases used in Experiment 1, using a logarithmic scale as explained before. Additionally, conditions were matched on familiarity and imageability ratings taken from the combined Bristol/G&L norms and also in word length (in letters) and Coltharts N measure of orthographic neighbourhood. Table 2.6 presents a summary of word characteristics for each condition and p values for the t-test between conditions for each of the relevant variables mentioned above. Appendix C shows a complete list of the items included and individual values for each variable.

Table 2.6: *Summary of item characteristics for Experiment 2.*

	AoA	IMG	FAM	LEN	N	Log WFGF	Log CelexF	Log BNCF
Early	221	499	513	5.31	2.71	1.11	1.16	1.08
Late	564	486	499	5.60	2.03	1.00	1.21	1 20
p	< .01	.61	.28	.30	.29	.29	.63	.28

Note: AoA = Age of Acquisition; IMG = Imageability; FAM = Familiarity; LEN = length in letters; N = orthographic neighbourhood; WFGF = word frequency according to the Educator's Word Frequency Guide; CelexF = word frequency according to CELEX; BNCF = word frequency according to the British National Corpus.

Results.

Mean reaction times, mean z-transformed reaction times and error rates for Experiment 2 are presented in Table 2.7. Appendix C presents detailed information on an item-by-item basis. Once again, significance tests were performed on the z-transformed reaction times and all analyses presented are by items. For lexical decision, when AoA was manipulated keeping word frequency constant, reaction times for early words (e.g. *elbow*) were also faster than for late-acquired words (e.g. *election*) ($F(1,68) = 9.27$; $p < .01$) with no trade-off in accuracy ($F(1,68) = 0.47$; $p = .50$). For word naming, and once again noting the caveats raised in Experiment 2, the pattern is similar: early-acquired words are read faster than late-acquired words ($F(1,68) = 4.58$; $p < .05$).

Table 2.7: *Summary of results for Experiment 2.*

	LD_RT	LD_Z	LD_Acc	NMG_RT	NMG_Z
EarlyAoA	614	-0.60	0.97	611	-0.52
LateAoA	649	-0.48	0.96	634	-0.43

Note: LD_RT = mean reaction times for lexical decision; LD_Z = standardized z-scores for lexical decision; LD_Acc = accuracy rate for lexical decision; NMG_RT = mean reaction times for word naming; NMG_Z = standardized z-scores for word naming.

Discussion

In summary, clear Frequency and AoA effects were found while the other variable was held constant. Each condition comprised a large number of items, and items in each condition were tightly matched on a variety of variables. These results confirm the findings of Morrison and Ellis (1995) and others about the existence of independent effects of word frequency and AoA, even when Zevin and Seidenberg's (2002) concerns about possible confounds are addressed.

2.4. Index of Norming Studies for Age of Acquisition, Imageability and Familiarity.

Section 2.2 introduced the Bristol Norms for Age of Acquisition, Imageability and Familiarity, which were collected in order to increase the availability of materials for experimental purposes. A variety of previous rating studies have collected similar information in different languages and under different conditions. As a complement to the Bristol Norms, this section presents an index of published rating studies for the same variables included in the Bristol Norms. This is an attempt to concentrate information into a usable reference guide that will help in the selection of stimuli for future research in the area. The index comprises studies with ratings for at least 100 items in one or more of the variables included in the Bristol Norms (i.e. AoA, imageability and familiarity). The studies are organized by language. Table 2.8 presents studies for which the ratings have been performed on pictures (or the names of the pictures), while Table 2.9 presents studies for which the ratings have been performed on words.

Table 2.8: *Ratings performed on pictures.*

Study	N	Var.	Other Variables	Comments
Chinese				
Chen, Yen, Tsai & Yeh (2001)	105	FAM	Naming accuracy, NamAg, subjective complexity, ImAg	S&V
Shu, Zhang, Li & Wang (1992)	260	FAM	NamAg, ImAg, VisComp	B/W pictures
Dutch				
Van Schagen et al. (1983)	260	FAM	NamAg, ImAg, VisComp	S&V
English				
Berman, Friedman, Hamberger, & Snodgrass (1989)	321	FAM	NamAg, VisComp	Children and adult. S&V plus others
Carroll & White (1973)	220	AoA		
Cycowicz, Friedman, Rothstein & Snodgrass (1997)	400	FAM	NamAg, VisComp	Children adults. Line drawings.
Masterson & Druks (1998)	266	AoA FAM IMG	NamAg	Ratings for printed labels of object & action pictures.
Morrison, Chappel & Ellis (1997)	297	AoA FAM IMG	Rated Freq, PicNamAg, NamAg	Objective & Rated AoA
Snodgrass & Vanderwart (1980)	260	FAM	NamAg, ImAg, VisComp	S&V
Snodgrass & Yuditsky (1996)	250	AoA		S&V
Winters, Winter & Burger (1978)	456	AoA	Rating confidence	Colored pictures.
French				
Alario & Ferrand (1999)	400	AoA FAM	NamAg, ImAg, VisComp, ImVar	Word AoA, Picture FAM. Line drawings.
Bonin, Peereman, Malardier, Meot & Chalard (2003)	299	AoA FAM	NamAg, ImAg, VisComp, ImVar	B/W drawings

Study	N	Var.	Other Variables	Comments
Bonin, Boyer, Macot, Fayol & Droit (2004)	142	AoA FAM IMG	NamAg, ImAg, VisComp, duration of actions	Action photographs
Chalard, Bonin, Meot, Boyer & Fayol (2003)	230	AoA		Objective AoA
Icelandic				
Pind, Jonsdottir, Gossurardottir & Jonsson (2000)	260	AoA FAM IMG	NamAg, ImAg	Objective & Rated AoA. S&V.
Italian				
Dell'Acqua, Lotto & Job (2000)	266	AoA FAM	Within-category typicality, NamAg	Italian PD/DPSS set. Line drawings.
Nisi, Longoni, Snodgrass (2000)	260	AoA FAM	NamAg	S&V
Portuguese				
Pompeia, Miranda & Bueno (2003)	400	FAM	VisComp, NamAg	Brazilian. S&V.
Spanish				
Cuetos & Alija (2003)	100	AoA FAM IMG	NamAg, VisComp	Action pictures
Cuetos, Ellis & Alvarez (1999)	140	AoA FAM	VisComp, ImAg, NamAg	S&V
Perez & Navalon (2003)	290	FAM	NamAg, ImAg, VisComp, ImVar	
Pineiro, Manzano & Reigosa (1999)	257	AoA FAM	Naming dispersion, NaMag, complexity, typicality	Cuban children. S&V.
Sanfeliu & Fernandez (1996)	254	FAM	NamAg, ImAg, VisComp	S&V
Aveleyra, Gomez, Ostrosky & Rigalt (1996)	260	FAM		Mexican Spanish. S&V
Perez & Navalon (in press)	175	FAM		Objective & estimated AoA

Note: N = number of raters; AoA = Age of Acquisition; IMG = Imageability; FAM = Familiarity; NamAg = Name Agreement; ImAg = Image agreement; VisComp = Visual Complexity; Freq = Frequency; ImVar = Image Variability; S&V = Snodgrass and Vanderwart's (1980) set of pictures; B/W = Black and White.

Table 2.9: *Ratings performed on words.*

Study	N	Var.	Other Variables	Comments
Dutch				
Ghyselinck, Custers & Brysbaert (2003)	2332	AoA		Words from 49 different semantic categories
Ghyselinck, DeMoor & Brysbaert (2000)	2816	AoA		4 & 5 letter nouns
Hermans & De Houwer (1994)	740	FAM	Affective familiarity	Nouns and personality-trait words
English				
Altarriba, Bauer, & Benvenuto (1999)	326	IMG	Conc, context availability, word associations	Abstract, concrete, and emotion words
Azuma (1996)	110	FAM	Meaning relatedness	Homographs
Benjafield & Muckenheim (1989)	1046	IMG FAM	Conc, goodness	
Berrian, Metzler, Kroll & Clark-Meyers (1979)	324	IMG	Ease of definition, animateness	Adjectives
Bird, Franklin & Howard (2001)	2645	AoA IMG		Disambiguated grammatical category. Young and old raters
Chiarello, Shears, & Lund (1999)	1197	IMG	Distributional typicality	Pure nouns, pure verbs, or words of balanced noun-verb usage
DiVesta & Walls (1970)	487	IMG	Emotionality, Conc	Ratings from 5th graders and undergraduates
Friendly, Franklin, Hoffman, & Rubin (1982)	1080	IMG	Conc, orthographic variables, grammatical usage	
Gilhooly & Hay (1977)	205	AoA IMG FAM		Anagram solutions. 5-letter words.
Gilhooly & Logie (1980a)	1944	AoA IMG FAM	Conc, Ambiguity	

Study	N	Var.	Other Variables	Comments
Gilhooly & Logie (1980b)	387	AoA IMG FAM	Conc	Ambiguous words
Kerr & Johnson (1991)	161	IMG FAM	Conc, Mean, imagery modality, word associations	Blind and seeing raters. Nouns
Morris & Reid (1972)	925	IMG		British and Canadian English
Paivio, Yuille & Madigan (1968)	925	IMG	Conc, Mean	Nouns, Canadian raters
Stratton, Jacobus & Brinley (1975)	543	AoA IMG FAM	Mean	4 & 5 letter words
Toglia & Battig (1978)	2627	IMG FAM	Conc, pleasantness, number of attributes or features, categorizability	
Walker (1970)	338	IMG		Nouns
French				
Bonin et al. (2003)	866	IMG	Conc, subjective, emotional valence	
Desrochers & Bergeron (2000)	1916	IMG	Subjective Freq	
Ferrand, Grainger & New (2003)	400	AoA		Monosyllabic words
German				
Hager & Asmuss-Kumke (1996)	855	FAM	Degrees of threat	Adjectives, verbs & nouns
Offe, Anneken & Kessler (1981)	234	IMG	Conc	Nouns
Italian				
Barca, Burani & Arduino (2002)	626	AoA IMG FAM	Conc	
Spanish				
Callejas, Correa, Lupianez & Tudela	612	FAM	Intra-categorical associative strength	Words From Six Semantic Categories
Perea (1993)	879	FAM		4-letter words

Study	N	Var.	Other Variables	Comments
Pineiro & Manzano (2000)	1259	AoA		Cuban children.
Sebastian, Marti, Carreiras y Cuetos (2000)	6286	IMG FAM	Conc	
Valle-Arroyo (1998)	4959	IMG		
Welsh and English				
Fear (1997)	705	AoA IMG FAM	Conc	
Catalan and Spanish				
Nacher, Gotor & Algarabel (1998)	1533	FAM	Conc	

Note: N = number of raters; AoA = Age of Acquisition; IMG = Imageability; FAM = Familiarity; Freq = Frequency; Conc = Concreteness.

Chapter 3. Age of Acquisition and Frequency Effects in Visual Word Recognition: Evidence from Expert Vocabularies.

Chapter 3

AGE OF ACQUISITION AND FREQUENCY EFFECTS IN VISUAL WORD RECOGNITION: EVIDENCE FROM EXPERT VOCABULARIES[4]

3.1. Introduction

As mentioned in Chapter 1, there is a controversy about the relative importance of frequency effects and AoA in visual word recognition. One possible explanation that has been proposed in the literature is that AoA effects could be reduced to cumulative frequency; that is, the total number of times that a word has been encountered (e.g., Lewis, 1999a, 1999b; Lewis et al., 2001; Zevin & Seidenberg, 2002). According to this account, the specific time at which these exposures occurred is not relevant; all else being equal, a word with a low frequency that has been encountered once in a while for several years should have the same performance than a word that was learned much later but that is encountered much more often if the total number of exposures to each word is the same. Consider the words *bubble* and *organic* that have equal frequencies in the CELEX norms[5] (F = 15) but have different AoAs (4.26 and 12.92 years, respectively). If reaction times to *bubble* were shorter it might appear to be an AoA effect given that the items were matched on CELEX frequency. However, given that *bubble* is encountered earlier in life, its cumulative frequency is higher than *organic*, in which case the reaction time difference might reflect cumulative frequency. It is therefore important to determine whether AoA effects can be attributed to cumulative frequency or not. Despite the large number of studies that have attempted to determine the relative importance of these two factors, a definitive answer has proven elusive, with some

[4] Adapted from Stadthagen-Gonzalez, Bowers and Damian (2004).

[5] All CELEX and Expert Frequencies are given in counts per million through the thesis. All CELEX frequencies are taken from the ECT database.

evidence supporting the frequency account (Zevin & Seidenberg, 2002, 2004) and other reports supporting the existence of independent AoA effects (e.g., Ghyselinck et al., 2004).

Ellis and Lambon-Ralph (2000) ran a series of neural-network simulations in which they showed an advantage for patterns that were presented to the network earlier than others. They attributed this advantage for early-trained patterns to a loss of plasticity in the network; that is, a reduction in the network's ability to modify its weights to adjust to later acquired patterns. In two of their simulations, "Extended Training on Early and Late Sets Under Conditions of Cumulative, Interleaved Training" (Simulation 3) and "Age of Acquisition Does Not Reduce to Simple Differences in the Cumulative Frequency of Early and Late Patterns" (Simulation 4), Ellis and Lambon-Ralph explicitly addressed the issue of whether AoA effects could be reduced to cumulative frequency in the model. In Simulation 3 they showed that the advantage for early-acquired patterns was preserved even after extensive interleaved training on both early and late sets of patterns in such a way that the difference in cumulative frequency between them became negligible. In Simulation 4 the network was first trained on an early set of patterns for 1,000 learning cycles, at which point the late set was introduced and both sets continued to be presented to the model for a further 1,000 cycles. However, this late set was presented to the network twice on each cycle, so that at the end of the training phase each set had been presented to the model exactly 2,000 times. The authors found that the advantage for the early-acquired set was present even under these circumstances. These results show that AoA effects cannot always be reduced to cumulative frequency in a PDP model.

Furthermore, in simulation 13, "Very Late Acquisition of Vocabulary", Ellis and Lambon-Ralph (2000) explored what happened when new patterns where presented to the model very late in the training phase, once the network has achieved a stable state. These conditions are akin to the introduction of new words in the vocabulary late in life due to technological advances or jargon associated to a particular adult occupation; the authors give the example of the word *email*, which is late acquired for most people over 20 years of age. Ellis and Lambon-Ralph found that it was very difficult for these very late acquired items to match the performance of earlier ones; indeed the very late items were never able to catch up with early acquired items, even when their total number of presentations was several orders of magnitude larger than for early acquired items. The authors acknowledge that this difficulty in acquiring very late items may be somewhat exaggerated in this simulation and propose that some degree of weight decay should improve the receptivity of the network to new information and help generalize old learning to the late patterns (Plaut, 1997; Plaut et al., 1996). Nevertheless this simulation predicts that late acquired items' representation in memory will not be as well established as early-acquired items, and that frequent exposure to them at a later age will have a limited impact on performance.

Zevin and Seidenberg (2004), however, pointed out that the random binary strings used in Ellis and Lambon-Ralph's simulations were not realistic representations of words, particularly with respect to their lack of orthographic structure and the fact that, despite a degree of consistency between input and output patterns, there was little overlap amongst the patterns used; in other words, their model did not include orthotactic and phonotactic information that is present in real languages. In most languages, orthographic and phonological representations of words observe a certain degree of regularity that allows for some of the learning attained by early-acquired items to be transferred to later-acquired items. When the authors used a model that included orthotactics and phonotactics and systematic mappings between them, the advantage for early-acquired words disappears because late-acquired words also partially benefit from the same modifications in weights during training of earlier acquired word. For example, the orthographic forms *prince* and *print* map onto similar output patterns, the phonological forms /prIns/ and /prInt/, respectively. Under these conditions, early learning (*prince* → /prIns/) contributes to late learning (*print* → /prInt/), eliminating any AoA differences. Zevin and Seidenberg (2004) proposed that the results from Ellis and Lambon-Ralph's (2000) simulations were only valid when there was no systematic mapping between input and output patterns, a condition that is true for tasks such as picture naming, but not for word recognition in English and other relatively shallow languages.

According to Zevin and Seidenberg, their simulations lend support to the idea that AoA effects can indeed be reduced to cumulative frequency. Zevin and Seidenberg also pointed out that methodological problems related to the way in which frequency and AoA are estimated compromise the validity AoA effects found in previous studies (Gerhand & Barry, 1999a, 1999b; Lund & Burgess, 1996; Monaghan & Ellis, 2002; Morrison & Ellis, 1995; Turner et al., 1998) that used the Kucera and Francis (1967) norms. In particular, Zevin and Seidenberg showed that when better norms were used, words in the early-acquired conditions were often higher in frequency than late-acquired items, suggesting that their advantage was the product of frequency differences rather than AoA.

Of course it is not possible to eliminate all measurement errors associated with frequency corpora or obtain an exact measure of cumulative frequency for each participant. Nevertheless, a strong test of AoA vs. cumulative frequency could be made if the estimated separation between conditions in a factorial design was large enough to render the inevitable measurement errors irrelevant. In order to achieve this, and inspired by Ellis and Lambon-Ralph's (2000) simulation, *Expert Frequency Databases* were created with words acquired late in life but that have a very high frequency for a specific population. Using these databases along with CELEX it was possible to select words with differences in AoA and frequency that are much larger than past studies (see Tables 3.1, 3.5 and 3.7). If, as proposed by Zevin and Seidenberg (2002), only cumulative frequency plays a role in visual word recognition, one would expect late-acquired words with an overwhelmingly larger frequency count to have a substantial advantage over early-acquired words with a much lower cumulative frequency in both the lexical decision task and naming. It is important to note here that, although Zevin and Seidenberg's simulations are only meant to model word naming, the authors extend their methodological and theoretical considerations to other tasks such as lexical decision and word recognition in general. It is therefore important to test their predictions in lexical decision as well as word naming.

3.2. Experiment 3. Lexical Decision with Chemistry and Psychology Words.

In this experiment we used a lexical decision task (LDT) with a full factorial design to evaluate the relative importance of frequency and AoA in word processing. Psychologists and Chemists were chosen for this task because both disciplines have developed their own jargon that is both specific and independent of each other. The advantages of this approach are threefold. First, it facilitated the identification of High-Frequency/Late-Acquired words (which are normally very difficult to find), making it possible to run a full factorial design with words in all AoA and frequency conditions. Second, it permitted a controlled look at the effects of frequency for late acquired words. Both Chemistry and Psychology "expert" words used were learned relatively late in life by both groups of professionals (and at roughly the same age), but Chemistry words were very high frequency for Chemists and very low frequency for Psychologists. The opposite is true for Psychology words. By manipulating the subject groups while keeping the test items constant, words were matched on all other possible variables except for frequency. Finally the differences between conditions for both variables were much larger than past studies. For instance, the difference in AoA between late acquired/high frequency words and early acquired/low frequency words was 8.95 years, compared with a range of 4.22 years in the words used in Gerhand and Barry (1998). Frequency differences were also very large (at least a proportion of 16:1 between high- and low-frequency items, but typically much larger), so that there is little chance for measurement errors to produce an overlap between conditions. This allows for a clearer distinction between the groups of words belonging to each condition.

Of particular interest to the present study was to determine whether, as predicted by some of the simulations of Ellis and Lambon-Ralph (2000), early acquired words maintain an advantage over late acquired words even when the later ones enjoy a much higher frequency count.

3.2.1. Method

Participants

Twelve PhDs in experimental psychology and twelve in chemistry were tested. Participants were faculty members and post-doctoral researchers at the departments of psychology and chemistry at the Universities of Bristol and Cardiff. All were British English native speakers and had normal or corrected-to-normal vision and took part in the study as volunteers or in exchange for a movie ticket. The experimental psychologists had an average age of 41.4 years (range: 31-68 years) and had received their PhDs an average of 13.6 years before the experiment (range: 2-29 years). The chemists had an average age of 36.7 years (range: 27-45 years) and had received their PhDs an average of 10.2 years before the experiment took place (range: 2-19 years). All participants provided a rating of "Academic Reading"; a self-estimate of the proportion of all their reading belonged to their own area of expertise. On average, academic reading constituted 52% of the psychologist's and 48% of the chemist's reading material (see Table 3.2).

Design and Materials.

<u>Expert Frequency Databases.</u>

In order to obtain an estimate of frequency for words that are largely specific to certain fields of study (i.e., jargon), an *Experimental Psychology Expert Frequency Database* and a *Chemistry Expert Frequency Database* were created using recent electronic editions of three journals from each discipline. These journals were selected because they are highly specialized in a particular area and yet general enough to be read by most researchers in their discipline. They were also considered to be representative of each discipline by a panel of 3 university lecturers from each field that were consulted about the matter. Finally, a considerable number of electronic issues from each of those journals were available online at the time the databases were assembled. All electronically available articles from each journal were downloaded as PDF files and a UNIX script was used to count the number of instances of each type that was present on them. Each of the six lists obtained was then visually inspected in order to eliminate spurious entries such as abbreviations, misspellings and non-alphabetical characters. The *Experimental Psychology Expert Frequency Database* contains a total of 3,193,794 tokens, of which the journal *Cognition* contributed 846,213 tokens; the journal *Cognitive Psychology*, 1,574,683 tokens; and the *Quarterly Journal of Experimental Psychology*, 772,898 tokens. The *Chemistry Expert Frequency Database* is comprised of a total of 2,999,292 tokens, of which the journal *Chemistry* contributed 1,151,593 tokens; the journal *Chemical Society Reviews*, 553,057 tokens; and the *New Journal of Chemistry*, 1,294,642 tokens. The frequency count for each type was then normalized to reflect its relative frequency in a sample of one million tokens, consistent with other frequency corpora (e.g., CELEX). The "expert frequency" estimate for words was the average of the frequency (in counts per million) for each word from all three journals within each discipline. A word was classified as "high expert frequency" if it fulfilled two conditions:

a) The word had a frequency of more than 100 in each of the three journals. This condition was set up so that high expert frequency words were representative of the entire field, and not just the result of a potential skew of one of the journals towards a certain group of words.

b) The word had a frequency of 15 or less on the *Expert Frequency Database* of the other discipline *and* in CELEX. This condition limited the possibility that high expert frequency estimates received significant contributions from the general language and so that they could be used in the frequency manipulation described below.

A word was classified as low-frequency if its expert *and* CELEX frequencies were 15 or less.

<u>Ratings for AoA</u>

AoA estimates (in years) were provided by twelve psychology and twelve chemistry graduate students at Bristol University. They provided the ratings in exchange for a chocolate bar. AoA ratings have been shown to be valid and reliable estimates of the real age at which words are acquired. All items in the experiment were randomised and presented on the same session along with 32 other items that were known to span a good range of AoA. Participants were asked to indicate the age at which they would have understood the spoken form of each word. The exact wording of the instructions provided to participants can be found in Appendix A.

A word was classified as "late-acquired" if its rating was above 10 years and "early-acquired" if it was below 6 years.

Stimuli

Four groups of 16 words each were constructed: (1) Psychology words: late AoA and high psychology expert frequency (e.g., *cognition*); (2) Chemistry words: late AoA and high chemistry expert frequency (e.g., *electron*) (3) Early/LoF words: low CELEX frequency and early AoA (e.g., *dragon*); (4) Early/HiF words: high CELEX frequency and early AoA (e.g., *smile*). The words in conditions 1 and 2 served as Late/HiF items when read by the relevant expert (e.g., psychologist reading *cognition*), and Late/LoF items otherwise (e.g., chemist reading *cognition*). Words in each group were controlled for length and there were no significant differences in orthographic neighbourhood size (all F values < 1.26, p values $> .27$). Table 3.1 presents a summary of item attributes and Appendix D shows an item-by-item list of those details. Frequency counts from the Educator's Word Frequency Guide (Zeno et al., 1995) are also presented.

A set of non-words was created for the lexical decision task by changing a letter from additional words taken from the Expert Frequency and CELEX Databases (e.g., *castoral, mirlion*). A complete list of the nonwords used is included in Appendix D. All non-words

Table 3.1: *Summary of item characteristics for Experiment 3.*

	PsyF	ChemF	CelexF	WFGF	AoA	N	Len
Psychology	398	2.0	3.6	2.4	14.0	1.6	7.4
Chemistry	2.3	680	5.9	13.4	13.2	0.7	7.5
Early/LoF	-	-	5.3	8.6	4.7	1.3	6.8
Early/HiF	-	-	83.5	83.4	4.4	1.8	7.2

Note: PsyF = Psychology Expert Frequency; ChemF = Chemistry Expert Frequency; CelexF = CELEX frequency; WFGF = Educator's Word Frequency Guide Frequency; AoA = Age of Acquisition; N = Orthographic Neighbourhood; Len = Length (in letters).

were pronounceable, and none were homophonic to real words. Words and non-words were matched in length and there were no significant differences (all F values < 1.26, p values > .27) in length-sensitive token bigram frequency between words and non-words according to Davis (2005). Bigram frequency has been used as a measure of how well a letter string conforms to the English spelling patterns (Novick & Sherman, 2004).

<u>Estimated Cumulative Frequency of items.</u>

We estimated that the cumulative frequency for Late/HiF words was at least 15 times larger than Early/LoF words and very similar to the cumulative frequency of the Late/HiF items (see Table 3.2). The Estimated Relative Cumulative Frequency is a rough estimate of the proportion of cumulative use of a word given in tokens per million. The estimate for both early conditions was obtained by multiplying the average CELEX frequency of these words by the years of use (that is, the age of the participants minus the AoA of the words). For late words, the values from the Expert Frequency Databases were multiplied by the number of years since the participant obtained a PhD and by the mean percentage of academic reading for each group. The usage of these words before PhD was not included, although if taken into account this would only strengthen the point illustrated here.

Even if this calculation were overestimated by a factor of 10, words classified as late-acquired would still have a larger cumulative frequency than early words.

Procedure

The LDT was self-paced; participants were instructed to press the space bar to initiate each trial, and to press the right Shift key if the letter string displayed on the screen was a word, and the left Shift key otherwise. The need for both speed and accuracy was stressed at the beginning of the experiment. After pressing the space bar, a plus sign appeared on the centre

Table 3.2: *Relative ratio of cumulative frequency for all experiments in this chapter.*

	Age (years)	Years with PhD	Academic Reading	Ratio of Frequency	Cumulative
				Late/HiF to Early/LoF	Late/HiF to Early/HiF
Psychologists (Exp.3)	41.4	13.6	52%	15.3	0.9
Chemists (Exp.3)	36.7	10.2	47%	20.3	1.2
Geologists (Exp.4)	41.7	13.6	52%	13.3	0.2
Geologists (Exp.5)	44.1	14.5	63%	26.3	0.2

of the screen, it disappeared, and 500 milliseconds later a string of letters appeared centered around the position of the prompt. Participants received feedback concerning their accuracy and speed immediately after each trial. If a response was not entered within 2000 milliseconds, that trial was terminated and the response registered as en error. Participants responded to a set of 16 practice items before the main experiment and they were allowed to repeat the practice if they deemed it necessary. Items were randomized separately for each participant. There were a total of 144 trials in the experiment (including practice items) and the session took no more than 15 minutes. Stimuli were presented in lower case Courier-New font, 10-point size, with black letters over white background. The experiment was run using DMDX (Forster & Forster, 2003).

3.2.2. Results and Discussion

Only the word data were analyzed, and the results are shown in Table 2. Scores more than 2.5 SD from the mean were removed (1.1%). The word *lexical* was excluded because its error rate for chemists was more than 50% and more than 2.5 SD above the mean. The word *chemistry* was also removed because its rated AoA (8.53 years) was too low for its assigned condition. Analyses were carried out both by subjects (F_1) and by items (F_2).

Collapsing across groups, Late/HiF words (e.g., a chemist's response to *carbon*) have an advantage over Late/LoF words (e.g., a psychologist's response to *carbon*) both for speed (difference of 72 ms; $F_1(1,23) = 42.02$, $p < .01$; $F_2(1,59) = 23.76$, $p < .01$) and errors(difference of 10.1% , $F_1(1,23) = 16.67, p < .01; F_2(1,59) = 12.04, p < .01$).

These results provide strong evidence for a frequency effect for late-acquired words that manifests itself on both reaction times and errors. Although the effects of frequency on word recognition have been reported in the past, the importance of the present results become apparent if one takes into account that high and low-frequency words were matched on all other possible variables (they were the same words), and only the frequency of exposure to the words was varied (by manipulating the population); the words that are low frequency for one of the subject groups are exactly the same words that are high frequency for the other group and vice-versa (e.g., *cognition* is high frequency for psychologists but low frequency for chemists, while *carbon* is high frequency for chemists but low frequency for psychologists). Variables that could be considered as potential contributors to lexical

Table 3.3: *Summary of results for Experiment 3.*

	RT (ms)	% Error
Late/HiF	581	3.8
Late/LoF	653	13.9
Early/LoF	580	4.7
Early/HiF	532	2.4

decision reaction times such as word length and orthographic neighbourhood (e.g., Morrison & Ellis, 2000) are all controlled for. Most importantly for the purpose of the present study, AoA can also be completely controlled across frequency conditions. These results are consistent with the findings obtained by Gardner, Rothkopf, Lapan and Lafferty (1987) who observed that nurses responded faster to medical terms that lawyers and engineers.

The presence of a strong frequency effect while holding AoA ratings for each word constant is problematic for accounts of word recognition that attribute most frequency effects to confounded AoA effects (e.g., Carroll & White, 1973; Morrison & Ellis, 1995).

With regards to AoA, Late/HiF words (e.g., a chemist's response to *carbon*) did not have a significant advantage over Early/LoF words (e.g., a chemist's response to *dragon*) in RTs (difference of 1 ms, $F_1(1,23) < 1$; $F_2(1,61) < 1$) nor errors (difference of 1%, $F_1(1,23) < 1$; $F_2(1,61) < 1$). These results suggest that AoA is also a significant factor in word recognition, since a pure frequency account would predict much better performance for Late/HiF words given the extreme cumulative frequency differences between conditions. The lack of advantage for Late/HiF over Early/LoF words cannot be attributed to a floor effect since performance was better for Early/HiF (532ms; $F_1(1,23) = 55.57, p < .01$; $F_2(1,61) = 21.02, p < .01$; 2.4% errors, $F_1(1,23) < 1$; $F_2(1,61) = 1.38, p = .24$). AoA effects are also supported by the finding that RTs were much reduced for the Early/HiF compared to Late/HiF items (difference of 49 ms; $F_1(1,23) = 55.57, p < .01$; $F_2(1,61) = 21.02, p < .01$; errors: difference of 1.4%; $F_1(1,30) = 1.67, p = .21$; $F_2(1,61) = 1.38, p = .24$) despite items having comparable cumulative frequency counts (a ratio between Early/HiF and Late/HiF of 0.9 for psychologists and 1.2 for chemists, see Table 3.2).

A possible criticism of this first study with regards to the AoA findings is that items were not matched on bigram frequency, a measure of orthographic redundancy that some authors (e.g., Anisfeld, 1964; e.g., Rice & Robinson, 1975) consider to be relevant in visual word recognition (for an opposing view, see Andrews, 1992).[6] More importantly, items were not matched on measures of semantic complexity, such as subjective ratings of concreteness or imageability, that may affect performance in the LDT (e.g., Kroll & Merves, 1986; Paivio et al., 1968). Twenty psychology students were asked to provide post-hoc ratings for the words in this experiment on a 7-point scale; the exact instructions are presented in Appendix A. As can be seen in Table 3.4, these ratings showed significant differences between conditions, with Psychology Words rated as the most abstract, followed by Chemistry Words, Early/HiF and Early/LoF words. More concrete words (e.g., *chair*) tend to be recognised quicker and better than more abstract words (e.g., *faith*), so it could be that the lack of advantage for late acquired words (i.e. psychology and chemistry words) could be attributed to the fact that they are more abstract than the Early/LoF words. These concerns were addressed in Experiments 4 and 5. Obtaining and effect of AoA even when items across conditions are matched on concreteness or imageability would further support the claim that AoA effects are real and independent from other lexical factors.

3.3. Experiment 4. Lexical Decision with Geology Words.

As mentioned before, in Experiment 3 psychology and chemistry words were found to be more abstract than the other conditions, so there is a possibility that the results obtained could be attributed to a confound between AoA and concreteness. In order to address this issue, in this experiment geology words were used for the Late/HiF condition. Geology was chosen because this discipline has developed its own characteristic jargon with words that tend to be more concrete than most other scientific disciplines.

[6] Bigram frequencies were introduced in the analysis at the explicit request of a reviewer for Stadthagen-Gonzalez et al. (2004).

Table 3.4: *Concreteness ratings and bigram frequencies for words in Experiment 3.*

	Concreteness	Bigram Frequency (tokens)	Bigram Frequency (types)
Psychology	2.86	871	84
Chemistry	4.30	535	38
Early/LoF	6.24	710	66
Early/HiF	4.45	885	55

3.3.1. Method

Participants

Twenty persons with PhDs in geology were tested. Participants were faculty members and post-doctoral researchers at the departments of geology at the Universities of Bristol, Cardiff, Imperial College, University College London and Birkbeck College (see Table 3.2). The geologists had an average age of 41.7 years (range: 27-59 years) and had received their PhDs an average of 13.6 years before the experiment took place (range: 2-31 years).

Design, Materials and Procedure.

Expert Frequency Database

A *Geology Expert Frequency Database* of approximately 3.8 million tokens was created using the same procedure as in the previous experiment. Three geology journals were picked using the same criteria as before: *Geology*, with 566,818 tokens; *The GSA Bulletin*, 1,800,352 with tokens; and *The Journal of Geology*, with 1,477,179 tokens.

AoA and Concreteness ratings

Twenty geology graduate students provided estimates for AoA (in years) and Concreteness (on a 7-point scale) (see Table 3.5). The instructions for rating concreteness were identical to the one used by Spreen and Schulz (1966p. 460). An additional set of 16 items with low concreteness (e.g., *luck*) was also included in order to encourage participants to use the entire range of the scale. All items were rated in the same session and randomised for each participant. Ratings were obtained from Geology graduate students because semantic attributes of specialized words are likely to differ for experts compared to persons from the general population.

Materials

Three conditions of 16 words each were used: (1) Geology (or Late/HiF): late AoA and high geology expert frequency (e.g., *basalt*); (2) Early/LoF: low CELEX frequency and early AoA (e.g., *dragon*); (3) Early/HiF: high CELEX frequency and early AoA (e.g., *water*). Words on the three lists were closely matched for length and concreteness and there were no significant differences in neighbourhood (F values < 2.1, p values $> .16$). Geology and Early/Low words were also controlled for length-sensitive positional bigram frequencies (both by tokens and by types; t values < 1.2, p values $> .2$; based on CELEX; Davis, 2005). Table 3.5 presents a summary of item attributes and Appendix E shows an item-by-item list of those details.

The cumulative frequency for the Late/HiF (geology) words was estimated to be 13 times larger than for Early/LoF words and calculated in the same way as for previous experiments (see Table 3.2).

The characteristics of the non-words and the testing procedures used were similar to the ones in Experiment 3. All non-words used are presented in Appendix E.

Table 3.5: *Summary of item characteristics for Experiment 4.*

	GeoF	CelexF	WFGF	AoA	N	Len	Conc	BigF (tokens)	BigF (types)
Geology	492	3.4	6.1	11.8	0.7	6.4	5.6	800	49
Early/LoF	-	6.6	12.6	4.3	1.8	6.5	5.7	831	49
Early/HiF	-	400	487	4.2	2.0	6.5	5.6	1578	65

Note: GeoF = Geology Expert Frequency; CelexF = CELEX frequency; WFGF = Educator's Word Frequency Guide Frequency; AoA = Age of Acquisition; N = Orthographic Neighbourhood; Len = Length (in letters); Conc = Rated Concreteness; BigF = Bigram Frequency.

3.3.2. Results and Discussion

The same criteria for detecting outliers as in Experiment 1 were applied both for individual RTs (2.8% were dropped) and errors (the word *rhyme* had an error rate of 35% and was eliminated from further analyses). Table 3.6 shows the results. There was no advantage for Late/HiF (e.g., *zircon*) over Early/LoF words (e.g., *dragon*) for speed (difference of 11 ms; $F_1(1,19) = 1.30$, $p = .27$; $F_2(1,29) < 1$) nor errors (difference of 2.6%; $F_1(1,19) = 1.63$, $p = .22$; $F_2(1,29) = 1.73$, $p = .20$).

It is interesting to note here that a large portion of the difference in errors rates between these two conditions comes from the word daffodil (20% error), which has an unusual

Table 3.6: *Summary of results for Experiment 4.*

	RT (ms)	% Error
Late/HiF (Geology)	553	3.1
Early/LoF	564	5.7
Early/HiF	530	1.9

spelling. If this words is eliminated from the analyses the numerical differences become smaller still, both for error rates (difference of 1.5%; $F_2(1,28) = 0.78$; $p = .38$) and RTs (difference of 5 ms; $F_2(1,28) = 0.27$; $p = .61$).

Performance for Early/HiF words was better than the other conditions (530ms, $F_1(1,19) = 7.01$, $p < .05$; $F_2(1,30) = 6.54$, $p < .05$; 1.9% errors $F_1(1,19) < 1$; $F_2(1,30) = 1.11$, $p = .30$) ruling out a floor effect. The important finding here is not that performance for the Early/LoF and Late/HiF items was equivalent, but rather, that performance for the Late/HiF items was not much better given their frequency counts. Indeed, based on the cumulative frequency hypothesis, RTs and error rates for the Late/HiF items should have been more similar to the Early/HiF items, which is not the case. This pattern of results replicates the findings of Experiment 3, even when concreteness was matched across the critical conditions.

3.4. Experiment 5. Word Naming with Geology Words.

Although Zevin and Seidenberg (2002) claimed that AoA does not affect lexical decisions (a claim challenged by Experiments 3 and 4), their simulations only dealt with word naming. Accordingly, it is important to assess AoA and frequency effects in this task as well. For this experiment, imageability was chosen as a measure of semantic complexity instead of concreteness. As mentioned in Chapter 2, these two measures are closely related and ratings for them are quite similar for most words. However, some studies suggest that imageability is a better predictor of performance than concreteness (Bruyer & Strypstein, 1985; Marcel & Patterson, 1978; Richardson, 1975).

One of the difficulties with word naming experiments is that phonetic characteristics of stimuli affect the measurement of response latencies with voice-triggers (Kessler et al., 2002); some initial phonemes (e.g., plosives) activate the voice trigger earlier than others (e.g. fricatives). Some researchers have used delayed word naming in order to overcome this problem, but Rastle and Davis (2002) showed that there were remnant frequency effects even after a delay, casting doubt over the validity of latencies calculated using this method. They recommend matching items across conditions on full word onset or initial vowel in order to eliminate biases in the measurement of onsets. Following this advise, items in Experiment 5 were carefully chosen so that they were matched on full onset, along with other relevant variables.

3.4.1. Method

Participants

Fifteen PhDs in geology were tested (see Table 3.2). Participants were faculty members and post-doctoral researchers at the departments of geology at the Universities of Bristol, Cardiff, Imperial College, University College London and Birkbeck College. The geologists had an average age of 44.1 years (range: 31-64 years) and had received their PhDs an average of 14.5 years before the experiment took place (range: 3-32 years).

Design and Materials.

<u>Imageability Ratings</u>

Items in these two conditions were also tightly controlled for AoA and imageability, with ratings coming from twelve geology graduate students (see Table 3.7). Imageability scores were made on a seven-point scale and the instructions were identical to those in Paivio, Yuille and Madigan (Paivio et al., 1968, p. 4). A total of 193 words were rated on the same session (37 geology words as well as 72 and 84 words that were presumed to be early/low and early/high, respectively). All items were rated in the same session and randomised.

Stimuli

A total of 48 words were classified into the same three conditions as in Experiment 2, but this time they were matched for full onset or initial vowel, as well as length and neighbourhood (all F values < 1). Geology and Early/LoF words were matched in length-sensitive positional bigram and trigram frequencies (both by tokens and by types; t values < 1.2, p values > .2). Table 3.7 presents a summary of item attributes and Appendix E shows an item-by-item list of those details.

The cumulative frequency for the Late/HiF (geology) words was estimated to be 26 times larger than for Early/LoF words and calculated in the same way as for previous experiments (see Table 3.2).

Apparatus

The experiment was run on a portable computer using DMDX software. Responses were captured using a Sennheiser m@b40 headset microphone and recorded directly into the computer's hard drive.

Table 3.7: *Summary of item characteristics for Experiment 5.*

	GeoF	CelexF	WFGF	AoA	N	Len	IMG	BigF (tokens)	BigF (types)
Geology	452	5.1	7.3	12.4	1.8	6.6	5.5	1063	67
Early/LoF	-	4.25	4.5	4.9	2.8	6.6	5.5	915	46
Early/HiF	-	435	436	4.7	2.2	6.0	3.4	1638	47

Note: GeoF = Geology Expert Frequency; CelexF = CELEX frequency; WFGF = Educator's Word Frequency Guide Frequency; AoA = Age of Acquisition; N = Orthographic Neighbourhood; Len = Length (in letters); IMG = Rated Imageability; BigF = Bigram Frequency.

Procedure

Participants initiated each trial by pressing the space bar, a fixation cross then appeared in the center of the screen for 500 ms and was replaced by a target word. Participants were asked to name words as quickly and accurately as possible. If a response was not entered within 2000 milliseconds, that trial was terminated and the response registered as en error. Sixteen practice trials preceded the experiment and critical items were randomized for each participant. There were a total of 64 trials in the experiment (including practice items) and the session took no more than 15 minutes. Visual characteristics of stimuli were identical to the previous experiments.

3.4.2. Results and Discussion

Non-speech signals (i.e., lip-pops, clicks and external noises) were manually removed from the sound files prior to analysis. There were too few mispronounced items (0.6%) to allow an error analysis. Responses more than 2.5SD from the mean were removed (4 from geology, 2 from Early/Low, and 4 from Early/High). Reaction times were calculated using Runword (Kello & Kawamoto, 1998). Table 3.8 presents a summary of the results. Once again there was no significant advantage for Late/HiF (geology) over Early/LoF words (difference of 5 ms; $F_1(1,14) = 1.01, p = .33$; $F_2(1,30) < 1$), and this is not due to a floor effect (difference of 15 ms between Early/HiF and Late/HiF; $F_1(1,14) = 15.96, p < .01$; $F_2(1,30) = 9.16, p < .01$).

Once again, the pattern of results indicates that items in the Early/LoF condition benefit from an AoA effect that offsets the advantage conferred to the Late/HiF words by their higher frequency. Consistently with Experiments 3 and 4, these results are difficult to accommodate by the cumulative frequency hypothesis, since it would predict a significant

Table 3.8: *Summary of results for Experiment 5.*

	RT (ms)
Late/HiF (Geology)	473
Early/LoF	478
Early/HiF	458

advantage for the Late/HiF condition, which had a cumulative frequency many times larger than the Early/LoF. The presence of an AoA effect in word naming is particularly problematic for Zevin and Seidenberg's (2002) connectionist model that predicted AoA effects only when the mapping between inputs and outputs was arbitrary, which is not the cease for orthography to phonology conversions in English.

3.5. General Discussion

The present experiments provide evidence for the independent contribution of AoA and cumulative frequency in word processing tasks when the usual confounds between these factors are removed. Evidence for AoA effects comes from the similar RTs obtained for

Late/HiF (expert vocabulary) items and Early/LoF words in lexical decision (Experiments 3-4) and naming (Experiment 5) tasks despite the fact that the cumulative frequency of the HiF words was over an order of magnitude greater. It is important to note that the present conclusions are not based on null effects. Indeed, these conclusions would not be altered conclusion if the small differences in performance between the critical conditions were significant. Rather, the important point is that the Late/HiF items should be *much* faster than the Early/LoF items, which was not the case. Critically, this outcome is obtained when the various criticisms of past studies raised by Zevin and Seidenberg (2002) do not apply; namely, that confounds produced by measurement errors in word frequency and other variables could yield spurious results. At the same time, the finding that Late/HiF words (e.g., a chemist reading *carbon*) were responded to more quickly than Late/LoF items (e.g., a psychologist reading *carbon*) when items were matched on all possible variables (they were the same words) shows that frequency also plays a role in LDT performance (Experiment 3).

The AoA findings pose a challenge to standard Parallel Distributed Processing (PDP) accounts of word naming. According to Zevin and Seidenberg (2002), the reason the PDP model (and humans) fail to show AoA effects in word naming (and lexical decision) is that there is a systematic relation between orthography and phonology, such that similar input patterns (e.g., the orthographic forms *prince* and *print*) map onto similar output patterns (e.g., the phonological forms /prIns/ and /prInt/, respectively). Under these conditions, early learning (*prince* → /prIns/) contributes to late learning (*print* → /prInt/), eliminating any AoA differences. Zevin and Seidengerg showed that, in their model, AoA effects were present only when the mappings between inputs and outputs were arbitrary. Although this analysis appears correct in the case of their model, it mischaracterizes human behaviour given that strong AoA effects are obtained in naming and lexical-decision tasks.

Zevin and Seidenberg (2002) also note that there is some behavioral evidence that AoA effects are larger in tasks involving arbitrary input-output mappings (e.g., picture naming) compared to systematic mappings (e.g., word naming), consistent with PDP accounts of AoA (also see Monaghan & Ellis, 2002). However it should be noted that this prediction is not unique to PDP models, and indeed, the dual-route model of reading makes the same prediction. The reasoning is straightforward: AoA is a lexical variable, and accordingly, AoA effects will be reduced to the extent that sub-lexical grapheme-phoneme correspondences contribute to performance. This is the case in word but not picture naming in the dual-route framework. What PDP models of word naming do uniquely predict is that AoA effects are eliminated when the input-output mappings are systematic, and this prediction is falsified in the present studies.

Based on these findings, we would suggest that AoA (and frequency) effects may reflect the structure of lexical-orthographic and lexical-phonological representations themselves. That is, both Early and HiF words may have "stronger" lexical representations that are more easily accessed, just as Morton (1979), McClelland and Rumelhart (1981), and Davis (1999) have argued in the case of frequency. On this latter approach, AoA effects should be observed even under conditions in which input-output mappings are systematic, as is the case of the correspondences between orthography and phonology in quasi-regular languages such as English. This idea is compatible with the growing network model (Steyvers & Tenenbaum, 2005) which proposes that AoA effects stem from a higher degree of centrality of earlier acquired words in the semantic network and can also account for word frequency effects. However the present results do not necessarily support the idea that semantics is the sole locus of AoA; Experiment 5 found AoA effects in word naming, and it is not clear whether semantics actually play a role in performance on this task. Clearly, future work is required in order to determine whether networks that learn lexical codes can account for the independent contribution of AoA and frequency.

Chapter 4. Does Spoken Frequency Affect Visual Word Recognition?

Chapter 4

DOES SPOKEN FREQUENCY AFFECT VISUAL WORD RECOGNITION?

4.1. Introduction

As noted in Chapter 1, word frequency effects are one of the most reliable predictors of reaction times in visual word recognition. Frequency effects have been shown in a variety of paradigms, and many theoretical and computational models have sought to explain these effects. Experiments on visual word identification typically only use estimates of written frequency, and either implicitly assume that spoken frequency does not have a significant effect on visual word recognition, or alternatively, that written frequency can be used as a valid estimate of total exposure to a word. This may have more to do with the availability of written corpus data as compared to spoken data than with a specific theoretical stance on the matter. It is not until relatively recently that spoken corpora large enough to provide a representative sample of spoken language use have been produced (e.g. The British National Corpus "Demographic" database with 4.6 million tokens).

Theoretically, several models make implicit assumptions about whether spoken frequency matters in visual word recognition. For example, in serial search models (e.g., Forster, 1976), access files are modality specific, so that the order of search in the bins of the orthographic access file during visual word recognition only depends on written word frequency. This feature is a direct consequence of the assumption of autonomy of information between levels of processing. In the case of the logogen model, detection thresholds were originally proposed to be affected by any exposure to a word, whether in spoken or written form (Morton, 1969). In later versions of the model (Morton, 1979), logogens became modality specific in order to accommodate the absence of cross-modal priming in word recognition tasks (Winnick & Daniel, 1970). Connectionist models do not make specific claims as to whether spoken frequency would influence visual word recognition, although it is conceptually feasible to include connections for this to take place. In the dual route cascaded model of visual word recognition, Coltheart, Rastle, Perry, Langdon, and Ziegler (2001) made a pragmatic decision to use only written frequency in defining the baseline level of activation for a unit in the lexicon, but they do not rule out the possibility of using spoken frequency as well in the future: "At present, a unit's [frequency scaling variable] value in the phonological lexicon is set to the same value as its corresponding written frequency. A possibility for future implementations of the DRC model would be to change this to spoken word frequency" (p. 216).

The underlying, and often implicit, assumption of most of these models and experiments designed to test frequency is that some property of a word's representation is affected by the frequency with which a word is encountered in print. However, an alternative possibility is that the relevant variable is not how often a word is encountered in print, but just how often it is encountered in total, independently of its modality, or even that spoken frequency is the relevant variable by itself.

However, it is not entirely obvious that written and spoken frequencies are interchangeable. In fact, there are important differences in the ways in which we write and speak; for example, a comparison between the Kucera and Francis (1967) written corpus and the Dahl (1979) spoken corpus cited in Fromkin and Rodman (1993) shows that the word "I" is ten times more frequent in spoken language than in written language. Fromkin and Rodman also note that profanities and "taboo" words are far more common in the spoken corpus than in the written one, while almost all the prepositions are more common in writing. Tryk (1968) noted that most sources of written English tend to respond to editorial and aesthetic constraints that make them different from conversational material, a fact that is reflected, for example, in the use of more synonyms in writing in order to avoid repetition (Dahl, 1979).

The magnitude of the correlation between written and spoken frequencies varies across corpora, but it is not as high as one might expect. Lee (2003) compared frequency counts from three written frequency databases (Carroll et al., 1971; Kucera & Francis, 1967; Zeno et al., 1995) to three spoken frequency databases (Brown, 1984; Dahl, 1979; Howes, 1966) and found correlations that ranged between .61 and .74 (mean $r = .67$). In order to confirm these correlations using larger and more recent frequency corpora, I carried out an analysis between the "Written" and "Demographic" (spoken) databases of the British National Corpus (BNC). For words with more than one count per million in both databases, the correlation is 0.64, which, although highly significant, is far from perfect. Of around 10,000 words included in this correlation, about 32% have written and spoken frequencies that differ from each other by more than 0.63 in the logarithmic scale, which corresponds to the linear difference between 15 and 65 counts per million, a difference that most researchers would take as large enough to separate two conditions in a factorial design. Of the words that fulfill this criterion, 27% had higher written frequencies, while 5% had higher spoken frequencies. Furthermore, Brown and Watson (1987) found independent effects of spoken and written frequency when word familiarity was taken as the dependent variable, which led them to conclude that both frequency measures are not redundant of one another.

As can be seen from these considerations, it is not necessarily safe to assume that written frequency can be equated with spoken frequency. In fact, given that there is some evidence that visual word recognition is at least partially mediated by phonology (for a review, see Frost, 1998), it is feasible, and even likely, that frequency effects should be at least partially attributed to how many times a word is heard. Even if one takes the view that the differences between the two forms of frequency are not especially large for most words, an important theoretical question remains: what is the extent and nature of the influence of spoken frequency in visual word recognition?

To the author's knowledge, the only study to have addressed this issue empirically is the one conducted by Ziegler, Tan, Perry and Montant (2000). These authors found effects of spoken frequency for Chinese characters in both character naming and lexical decision. In their first experiment, they manipulated spoken frequency by varying the number of homophones of target Chinese characters while matching written surface frequency constant across conditions. Spoken frequency counts are based on the sum of frequencies of all homophones to a word, so characters with homophone mates have a higher spoken frequency than characters with no homophone mates even when their surface frequencies are held constant. Ziegler et al. report that there was a significant advantage for characters with homophone mates over characters with no homophones, which indicated a facilitatory effect of spoken frequency above and beyond the effects of written frequency. As the authors point out, this experiment has two possible weak points: first, characters with no homophones are rare in Chinese, and this result could be due to intrinsic characteristics of those characters. Second, it has been shown that recognition is faster for polysemous than for non-polysemous words (for a review see Kawamoto, Ferrar, & Kello, 1994) and that a character's phonology could be simultaneously activating all the homophone's meanings.

Ziegler et al. (2000) addressed these concerns in their second experiment by controlling the number of homophones in each condition, holding surface frequency across conditions constant while varying spoken frequency. To illustrate this, consider that, in English, the words *thyme* and *flair* have very similar written frequencies and one homophone mate each, but thyme has a very high frequency homophone (*time*), whereas *flair* has a homophone that is not so frequent (*flare*). In this case the pair *thyme/time* would be included in the high frequency condition while the pair *flair/flare* would belong to the low frequency condition. This experiment replicated the results of the first one, showing a reliable advantage in identifying characters with higher spoken frequency even when written frequency was matched across conditions. Ziegler et al. concluded that finding pure spoken frequency effects in Chinese, an ideographic language in which the phonological route is not required to identify written words, provides strong evidence for the activation of phonology, even when phonological access is not required. Ziegler et al. argued that the conclusions derived from those results may extend to alphabetical languages, in which it seems more likely that phonology would be automatically accessed. The authors concluded that "phonological frequency uniquely contributes to the traditional frequency effect. Thus, researchers who rely on exclusively nonphonological mechanisms or representations to account for the frequency effect may be forced to reconsider some of their most fundamental assumptions" (p. 237).

However, there are important differences between alphabetic and ideographic languages with respect to the way in which they are read. One issue of concern is that Ziegler et al.'s (2000) results show a facilitatory effect of high-frequency homophones, which is the opposite of what has been observed in English (Pexman et al., 2001). Also, mean reaction times in both experiments were very slow (at least 720 ms for lexical decision) compared with typical reaction times in alphabetical languages such as English (around 600 ms or less). This may indicate that the word recognition process in ideographic languages does not follow the same time course as in alphabetical languages. In view of these issues, it is not completely clear that these results could be easily extrapolated to alphabetical languages.

The purpose of the present chapter is to explore whether spoken frequency influences performance in word recognition tasks in alphabetical languages and should, therefore, be included as a relevant factor in empirical studies and theoretical models on this field. In order to do this, it was considered appropriate to use a direct manipulation of frequency, without relying on homophone mates. Ziegler et al. (2000) suggest that the homophone manipulation they used may not be suitable to explore this issue in alphabetical languages because, in this situation, homophones are visually similar, and orthographic inhibition between similarly spelled words may cancel out the spoken frequency effects (Grainger, 1990; Stone, Vanhoy, & Van Orden, 1997; Ziegler, Montant, & Jacobs, 1997), while this is not a problem in Chinese since homophone characters are not orthographically similar.

The general idea of the present study was to use words with "unbalanced" spoken and written frequencies, that is, words with large frequency differences between modalities. For example, according to the British National Corpus, the word *hello* has a written frequency of only 17 counts per million but a spoken frequency of 394 counts per million; on the other hand the word *chapter* has a written frequency of 162 counts per million but a spoken frequency of only 11 counts per million. Items such as these were compared with two control sets of words that had similar spoken and written frequencies; one of those control sets was matched to spoken frequency of the "unbalanced" set and another matched to its written frequency. Of particular concern was to control for Age of Acquisition (AoA), a variable that has been shown to have an independent effect in word recognition performance (see chapter 3 of the present dissertation) and that was not controlled in Ziegler et al's (2000) experiments. The importance of controlling for AoA when studying spoken frequency effects becomes clear if one takes into account that the correlation between AoA and spoken frequency is 0.65, whereas the correlation between AoA and written frequency is 0.38. [7] The lexical decision task was chosen for this experiment because frequency effects have been

[7] This correlation was performed using AoA ratings for 3,055 words from the Combined Bristol/MRC Norms (See Chapter 2 of this thesis). Written frequencies were taken from the "Written" database of the British National Corpus while spoken frequencies were taken from the "Demographic" database of that Corpus.

noted to be larger for this task than for word naming (Forster & Chambers, 1973; Frederiksen & Kroll, 1976), and because of the difficulties associated with selecting appropriate stimuli in word naming experiments (Kessler et al., 2002; Rastle & Davis, 2002). Also, phonology is not explicitly required in this task, which makes it possible to evaluate the diverging predictions of the various models of word recognition mentioned in the introduction to this thesis. Experiment 6 will contrast words with high written frequency and low spoken frequency with the control sets, while Experiment 7 will do the same for words with high spoken frequency and low written frequency.

4.2. Experiment 6. Words with high written frequency and low spoken frequency.

4.2.1. Method

Participants.

Thirty undergraduate students from Bristol University participated in this experiment as a course requirement. All were native British English speakers and reported normal or corrected-to-normal vision.

Materials

In this part of the experiment, the critical set of items (henceforth called "WrittenHi") was comprised of words with higher written frequency and lower spoken frequency (e.g., *chapter*). Two control sets of words were also selected: one with high frequency (HiFreq; e.g., *house*) and one with low frequency (LoFreq; e.g., *handle*) in both modalities. There were 29 items in each of these three conditions.

Written frequencies for this experiment were taken from the "Written" database of the British National Corpus and the spoken frequencies were taken from the "Demographic" database of that Corpus, which corresponds to mostly conversational material. The Demographic database was chosen over the Context-Governed database because it was considered to be more representative of every-day experience of spoken words. Most of the context-governed corpus is formal in nature; much of the materials included there, such as newscasts and speeches, are written first and then read out loud, so the underlying structure of this database is akin to the written database.

Spoken frequency was calculated by adding the frequencies of all the homophones of the words used, but only a few of the selected words had homophones and, in general, this had little impact on the overall frequency counts.

The words in the WrittenHi set were chosen so that the difference between spoken and written frequencies was 0.63 in logarithmic scale[8] and at least 50 counts per million in a linear scale, with no overlap between the two conditions. In the two control sets, the logarithm of the spoken and written frequencies did not differ by more than 0.11 in logarithmic scale[9].

The written logarithmic frequency of the WrittenHi words was matched with the HiFreq logarithmic frequency (2.1 in both conditions; $t(58) = 0.83, p = .41$). The spoken logarithmic frequency of the WrittenHi words was matched with the LoFreq logarithmic frequency (1.1 in both conditions; $t(58) = 0.27, p = .79$). Words from the WrittenHi set were matched with each of the control sets for word length (number of letters), neighbourhood density, imageability, Age of Acquisition, and length-sensitive bigram frequency by tokens and by types as measures of orthographic similarity (all F values < 1.0, p values > 0.31). Neighbourhood density and bigram frequency values were obtained using N-Watch (Davis,

[8] For comparison purposes, 0.63 in a logarithmic scale is the equivalent of the difference between 15 and 65 counts per million on a linear scale.

[9] For comparison purposes, 0.11 in a logarithmic scale is the equivalent of the difference between 10 and 13 counts per million on a linear scale.

2005). Norms of imageabililty and Age of Acquisition were obtained from 16 Bristol University undergraduates who rated words on a 7-point scale using the same procedures as Paivio, Yuille and Madigan (1968) and Gilhooly and Logie (1980a), respectively. Tables 4.1a and 4.1b present a summary of characteristics for each condition, and details for all items used in this experiment are listed in Appendix G.

A set of pronounceable non-words was constructed by changing a letter from words within the same frequency range as each condition. Non-words and words were matched in length and there were no significant differences in their length-sensitive bigram frequency both by

Table 4.1: *Summary of characteristics and comparison between items for Experiment 6.*

	Log Wr_f	Log Sp_f	AoA	IMG	LEN	N	BF_TK	BF_TP
WrittenHi	2.09	1.11	292	429	5.83	3.52	1219	51
HiFreq	2.08	2.07	284	420	5.55	4.66	1414	44
LoFreq	1.13	1.13	304	463	5.90	3.59	1100	42
Diff. (WrittenHi - HiFreq)	0.01	0.96	8	9	0.28	1.14	195	7
p (WrittenHi - HiFreq)	.85	<.01	.65	.78	.50	.39	.41	.44
Diff. (WrittenHi - LoFreq)	0.96	0.02	12	34	0.07	0.07	119	9
p (WrittenHi - LoFreq)	<.01	.76	.34	.30	.87	.96	.57	.29

Note: Wr_f = written frequency; Sp_f = spoken frequency; AoA = Age of Acquisition; IMG = Imageability; LEN = length in letters; BF_TK = length-sensitive bigram frequency by tokens; BF_TP = length-sensitive bigram frequency by types; Diff(WrittenHi-HiFreq) = absolute value of the difference between the WrittenHi condition and the HiFreq condition for each variable; Diff(WrittenHi-LoFreq) = absolute value of the difference between the WrittenHi condition and the LoFreq condition for each variable.

tokens and by types (all t values < 1.02, p values > .31). A complete list of the non-words used is presented in Appendix G.

Procedure

Participants were instructed to press the right Shift key on the keyboard if the letter string displayed on the screen was a word, and the left Shift key otherwise. The need for both speed and accuracy was stressed at the beginning of the experiment. Items were randomized for each participant and presented continuously with two equally spaced breaks. At the beginning of each trial, a plus sign appeared on the centre of the screen, it disappeared, and 500 milliseconds later a string of letters appeared centered around the position of the prompt. Participants received feedback concerning their accuracy and speed immediately after each trial. If a response was not entered within 2000 milliseconds, that trial was terminated and the response registered as en error. Participants responded to the set of 16 practice items before the main experiment and they were allowed to repeat the practice if they deemed it necessary. There were a total of 103 trials in the experiment (including practice items) and the session took no more than 15 minutes. Stimuli were presented in lower case Courier-New font, 10-point size, with black letters over white background. The experiment was run using DMDX (Forster & Forster, 2003).

4.2.2. Results and Discussion

Only the word data were analyzed. Individual scores deviating more than 2.5 SD from an item's mean were removed from the sample (2.1% of all responses).

Analyses were carried out both by participants (t_1) and by items (t_2). Words with error rates of 20% or more were removed; there was one such item in the WrittenHi condition (*net*), none in the HiFreq condition and three in the LoFreq condition (*blouse, quote* and *socket*). After removing those items, the two control sets were still well matched to the critical set in all the relevant characteristics (all F values < 1.57, all p values > .22). It is important to note that similar results were obtained whether those items were included in the analyses or not.

Table 4.2 shows a summary of results for this experiment and Appendix G includes latency and error data for each item. Reaction times for items in the WrittenHi condition (e.g. *chapter*) were not significantly different from items in the HiFreq (e.g. *house*) condition (difference of 3 ms; $t_1(29) = 1.29, p = .21$; $t_2(55) = 0.86, p = .39$) while there was a significant difference with the items in the LoFreq (e.g. *handle*) condition (difference of 14; $t_1(29) = -3.80, p < .01$; $t_2(52) = -2.57, p < .05$). There was no significant difference in error rates between the WrittenHi and the HiFreq conditions (difference of 0.1%; $t_1(29) = 0.15, p = .88$; $t_2(52) = -1.46, p = .15$), nor between the WrittenHi and the LoFreq conditions (difference of 1%; $t_1(29) = -0.79, p = .43$; $t_2(52) = -1.46, p = .15$).

This pattern of results implies that the influence of the spoken frequency of items in the critical condition is, at best, very small. The next experiment uses the same method as this one to test words that have a high spoken frequency but a low written frequency.

Table 4.2: *Summary of results for Experiment 6.*

	RT (ms)	% Error
WrittenHi	481	5.4
HiFreq	478	5.3
LoFreq	495	6.4

4.3. Experiment 7. Words with high spoken frequency and low written frequency.

4.3.1. Method

Participants

Thirty participants with similar characteristics to the ones in Experiment 6 took part in this experiment

Materials

In this part of the experiment, the critical set of items (henceforth called "SpokenHi") was comprised of words with lower written frequency and higher spoken frequency (e.g., *hello*). Conversely to Experiment 6, the frequency of the HiFreq control set was matched to the spoken frequency of SpokenHi, while the frequency of the LoFreq control set was matched to its written frequency. The logarithm of the spoken frequency of the SpokenHi words was matched with the HiFreq logarithmic frequency (2.2 in both conditions; $t(58)=0.13, p=0.90$). The written logarithmic frequency of the SpokenHi words was matched with the LoFreq logarithmic frequency (1.1 and 1.2; $t(58)=0.58, p=0.56$). Items from each of the control sets were matched to the SpokenHi set on the same variables as in Experiment 6 (all F values < 0.73, p values > 0.47). Table 4.3 presents a summary of characteristics for each condition, and details for all items used in this experiment are listed in Appendix H.

The non-words used had similar characteristics to the ones used in Experiment 6 and were matched to the words in length and length-sensitive bigram frequency both by tokens and by types (all t values < 0.98, p values > .33). The full list of non-words is also presented in Appendix H.

Table 4.3: *Summary of characteristics and comparison between items for Experiment*

	Log Wr_f	Log Sp_f	AoA	IMG	LEN	N	BF_TK	BF_TP
SpokenHi	1.16	2.19	258	456	5.48	4.69	1378	51
HiFreq	2.19	2.19	257	459	5.38	4.93	1557	47
LoFreq	1.21	1.22	266	463	5.41	4.48	1181	145
Diff. (SpokenHi - HiFreq)	1.03	0.00	1	3	0.10	24	179	4
p (SpokenHi - HiFreq)	<.01	.97	.94	.95	.77	.85	.58	.74
Diff. (SpokenHi - LoFreq)	0.05	0.97	8	7	0.07	0.21	197	94
p (SpokenHi - LoFreq)	.52	<.01	.69	.86	.83	.87	.51	.65

Note: Wr_f = written frequency; Sp_f = spoken frequency; AoA = Age of Acquisition; IMG = Imageability; LEN = length in letters; BF_TK = length-sensitive bigram frequency by tokens; BF_TP = length-sensitive bigram frequency by types; Diff(SpokenHi-HiFreq) = absolute value of the difference between the SpokenHi condition and the HiFreq condition for each variable; Diff(SpokenHi-LoFreq) = absolute value of the difference between the SpokenHi condition and the LoFreq condition for each variable.

Procedure.

The procedure used was identical to the one in Experiment 6.

4.3.2. Results and Discussion.

Only the word data were analyzed. Individual scores more than 2.5 SD away from the mean (by items) were removed from the sample (2.4% of all responses). Words with error rates of 20% or more were removed; there were three such items in the SpokenHi condition (*chap, pence* and *quid*), none in the HiFreq condition and four in the LoFreq condition (*bake, dozen, crawl* and *muddle*). After removing those items, the two control sets were still well matched to the critical set in all the relevant characteristics (all F values < 1.09, all p values > .30). It is important to note that similar results were obtained whether those items were included in the analyses or not.

Table 4.4 shows a summary of results for this experiment and Appendix H includes latency and error data for each item. There was a significant difference in latencies between the SpokenHi (e.g. *hello*) and the HiFreq (e.g. *house*) conditions (difference of 16 ms; $t_1(29) = 4.6$, $p < .01$; $t_2(53) = 2.78$, $p < .01$), but no significant difference between the SpokenHi and the LoFreq (e.g. *handle*) conditions (difference of 3 ms; $t_1(29) = -0.72$, $p = .48$; $t_2(53) = -0.14$, $p = .89$). There was no significant difference in error rates between the SpokenHi and the HiFreq conditions in the analysis by participants (difference of 1.5%; $t_1(29) = 1.47$, $p = .15$), but this difference was significant in the analysis by items ($t_2(53) = 2.17$, $p < .05$). Note that this trend in consistent with the results for latencies (i.e. that there are differences between words in the SpokenHi and HiFreq conditions). There was no difference in error rates between the SpokenHi and the LoFreq conditions (difference of 0.2%; $t_1(29) = -0.13$, $p = .90$; $t_2(53) = 0.09$, $p = .93$).

Once again, there doesn't seem to be much influence of the spoken frequency of items in the critical condition.

Table 4.4: *Summary of results for Experiment 7.*

	RT (ms)	% Error
SpokenHi	475	5.1
HiFreq	459	3.6
LoFreq	478	5.3

4.4. General Discussion

In both experiments, performance on unbalanced-frequency words was similar to performance on the control set matched to their written frequency and significantly different from performance on the control set matched to their spoken frequency. Contrary to what Ziegler et al. (2000) predict, these results seem to indicate that there is little or no influence of spoken frequency on visual lexical decision, and that spoken frequency has little influence over the mechanism that underlies those effects, whatever it may be. This lends empirical support for models in which the influence of frequency on visual word identification is dependent only on written counts, such as the revised logogen model (Morton, 1979). On the other hand, these results call into question strong phonological accounts of reading (e.g., Frost, 1998) as they also imply that access to phonology is not mandatory for tasks that do not explicitly require it, at least for the stages of word recognition affected by word frequency. The present results also provide constraints for models that are agnostic on this issue (e.g., Coltheart et al., 2001).

As mentioned before, previous models have taken the position that word frequency effects rely on how often a word is encountered in total, independently of its modality (e.g. the original logogen model); in this case, the relevant predictor of performance would be the sum of both spoken and written frequencies or "net frequency". However, this doesn't seem to be the case; in part A of the experiment, there are significant differences in net frequency between the WrittenHi condition and the HiFreq condition, and yet performance on both sets of words is very similar. The same is true for the SpokenHi and LoFreq conditions of part B. If net frequency was the relevant variable, one would expect large differences in performance between the critical sets and those control sets on both cases.

In their study, Ziegler et al. (2000) found an influence of spoken frequency in Chinese and speculated that those results could be extrapolated to alphabetical languages, but this does not seem to be the case. There are several possible reasons for the discrepancies between the present results and Ziegler et al's results. As mentioned in the introduction to this chapter, they obtained a facilitatory effect of high-frequency homophones, which is the opposite of what has been observed in English (Pexman et al., 2001). This would seem to indicate that the processes that underlie lexical decision in Chinese are fundamentally different from the ones in alphabetic languages such as English (although note that the orthographic similarity of English homophones may be the cause).

Furthermore, Ziegler et al. did not control for Age of Acquisition, which as shown in the present thesis and elsewhere, could have an impact on their results. It would seem that, given the complexities of the Chinese script, with its many symbols and deep mapping between orthography and phonology, it would take many years to learn even a portion of the most commonly used characters. This fact leaves open the possibility of large differences in AoA amongst words. With respect to Ziegler et al.'s first experiment, it could be that earlier AoA words have more homophones. Furthermore, as shown in Chapter 2 of this thesis, AoA and spoken frequency are highly correlated, so by not controlling for AoA, Ziegler et al. have left open the possibility for a confound.

Another issue that could explain this difference in conclusions is the speed of processing during word recognition during each of the experiments. Reaction times for Chinese seemed to be much slower than the ones presented here; the fastest mean reaction times in Ziegler et al's (2000) lexical decision tasks were 740 ms for the first experiment and 592ms for the second experiment, both for their "high frequency" condition, while mean reaction times for even the low frequency conditions in the present study were below 500 ms. It is possible that longer overall reaction times provide more opportunity for phonology to influence lexical access. Pexman, Lupker, and Jared (2001) propose that lexical decision is based primarily on orthographic representations, but with feedback connections from phonology to orthography in a highly interactive system. In this type of model, orthographic representations are activated first when a word is presented, with activation of phonology following quickly after that. The level of activation of orthographic units is then modified by feedback connections from phonology to orthography, but the decision is based mainly at the orthographic level. The level of feedback activation sent from phonology to orthography (and therefore the influence of phonology over the whole process) would be proportional to the time elapsed before a decision is made. As a consequence of this property, this model would predict lessened phonological effects under circumstances in which responses are achieved quickly, while the role of phonology would increase when responses were slower. Feedback connections consistent with this account can be found in both distributed processing models (e.g., Harm & Seidenberg, 1999) and dual route models (Coltheart et al., 2001). It is therefore possible that the spoken frequency effects obtained by Ziegler et al. reflect slower processing speeds in lexical decision in Chinese, maybe due to intrinsic characteristics of the Chinese language, such as the large number of characters that need to be mastered in order to read it.

The results presented here also have some implications on the debate about the extent of the interactivity between the orthographic and phonological systems in word recognition. The results obtained in these two experiments are by themselves not enough to settle this issue, but they seem to provide a certain amount of constraint on the extent of the level of interactivity between the two systems; that is, that the level of interactivity is not so profound that the phonological frequency feeds back into visual word identification.

Chapter 5. Summary and Conclusion.

Chapter 5

SUMMARY AND CONCLUSION

Both AoA and word frequency effects play a central role in the development of theories and models of visual word recognition and a better characterization of them is of great importance in the elaboration of such theories and models. The main purpose of the present dissertation was to advance our knowledge of these two effects (AoA and word frequency), firstly by providing evidence that both factors are relevant and have independent roles in visual word recognition, and secondly, by providing a better characterization of each effect. In the case of AoA, the discussion was centered around the issue of whether AoA effects can be attributed to cumulative frequency or not. In the case of word frequency, the discussion was centered on the degree to which spoken word frequency affects visual word recognition. Throughout the development of these main topics, other important issues related to AoA and frequency were considered, such as the relationship between these two factors and other lexical variables, and the validity of word familiarity as a good representation of word frequency.

The first chapter provided an overview of the different hypotheses that have sought to explain word frequency and Age of Acquisition effects, as well as a summary of the techniques used to measure those effects. Chapter 2 presented the *Bristol Norms for Age of Acquisition, Imageability and Familiarity*, a large set of ratings for more than 1,500 words that will hopefully prove to be a valuable resource in the selection of experimental materials for future study in the field. The ratings included in the *Bristol Norms* were then used to replicate seminal findings concerning AoA and word frequency effects while addressing some methodological concerns that have been mentioned in the literature. It was shown that each of the two variables have a significant effect on word recognition performance when words on each condition were matched on the other variable. The norms were also used to explore whether familiarity measures provide a valid representation of word frequency. It was found that subjective familiarity appears to be inferior to objective frequency counts as a measure of the frequency with which words are encountered in print, although it may offer a reasonably good measure of the frequency with which words are encountered in speech. To the extent that subjective familiarity taps into something beyond frequency, it appears to offer no advantage over more clearly defined measures such as objective frequency or AoA. On this basis, it seems pertinent to question the usefulness of this variable in psycholinguistic research. Additionally, this chapter presents an index of published studies that include norms for AoA, Imageability and Familiarity. This index might be of help in identifying sources for experimental materials.

Chapter 3 presented a series of experiments that explored whether AoA and frequency have independent effects in visual word recognition, and particularly, whether AoA effects can be reduced to cumulative frequency effects or not. This was done using "expert vocabularies", that is, words that were learned quite late in life but have very high frequency for a specific set of people (i.e., psychologists, chemists or geologists). The experiments included in this chapter provide evidence for the independent contribution of AoA and cumulative frequency in word processing tasks when the usual confounds between these factors are removed. Although the effects of frequency on word recognition have been reported in the past, the importance of the present results become apparent if one takes into account that high and low-frequency words were matched on all other possible variables (they were the same words), and only the frequency of exposure to the words was varied (by manipulating the population). Variables that could be considered as potential contributors to lexical decision reaction times are all controlled for and, most importantly for the purpose of the present thesis, AoA can also be completely controlled across frequency conditions. The results also suggest that AoA is a significant factor in word recognition: it was found that late-acquired, high-frequency words did not show a better performance than early-acquired, low-frequency words, despite the fact that the cumulative frequency of the late acquired words was over an order of magnitude greater. The same pattern of results was obtained in two lexical decision tasks (Experiments 3-4) and a naming task

(Experiment 5). These results are difficult to accommodate within the framework of the cumulative frequency account since a pure frequency account would predict much better performance for late-acquired, high frequency words given the extreme cumulative frequency differences between conditions

Chapter 4 considered whether spoken word frequency has an effect on visual word recognition or not. Words with "unbalanced" frequencies (i.e., high spoken frequency but low written frequency or vice versa) were compared with two control sets of words that had similar spoken and written frequencies; one of those control sets was matched to spoken frequency of the "unbalanced" set and another matched to its written frequency. In both experiments, performance on unbalanced-frequency words was similar to performance on the control set matched to their written frequency and significantly different from performance on the control set matched to their spoken frequency. These results seem to indicate that there is little or no influence of spoken frequency on visual lexical decision, and that spoken frequency has little influence over the mechanism that underlies those effects, whatever it may be.

This thesis has advanced knowledge in the area at three different levels: methodological, pragmatic and theoretical. Methodologically, it presents the *Bristol Norms for Age of Acquisition, Imageability and Familiarity,* which at 1,526 words is one of the biggest sets of norms for those variables to-date. One of the stumbling blocks in the research of AoA and frequency effects has been the lack of databases large enough to allow the selection of adequate stimuli sets while controlling for other relevant variables, so these norms by themselves constitute an important contribution to the field, since they significantly increase the offer of rated words that can be used in experimental designs. However, two characteristics of these norms make them even more useful: firstly, they were designed to be compatible with the Gilhooly and Logie (1980a) norms; when merged, the *Combined Bristol/G&L Norms* provided ratings for 3,394 words, by far the largest norming database for those three variables. Secondly, the *Bristol Norms* were collected using a modular approach that allows the incremental increase in the number of words added to the database, which allows for a continuous growth of the norms. Also included in this thesis is an index of published norming studies with ratings for the variables included in the Bristol Norms (i.e. AoA, imageability and familiarity). This index should be valuable to researchers in the field since it condenses in one place references to information that has been scattered across a variety of sources.

A further methodological contribution from this thesis is the use of "Expert Vocabularies" in the study of lexical variables. A precursor of this approach is the study carried out by Gardner, Rothkopf, Lapan and Lafferty (1987) who observed that nurses responded faster to medical terms that lawyers and engineers. However, those authors used a descriptive approach, without attempting to quantify differences in the frequency of use for words in different disciplines. Until a few years ago, the task of assembling frequency counts was laborious and time consuming, so that efforts were concentrated on general-use databases that attempted to represent general language use. However, the availability of representative samples of electronic texts for specific disciplines and the possibility of performing the word count using computer programs, allows for the elaboration of custom made frequency databases so that interesting differences in vocabulary use can be exploited to explore general issues of language processing, as was done in Chapter 3 of this thesis.

Pragmatically, this thesis sheds some light into which are the relevant variables that should be controlled for in visual word recognition experiments. It provides evidence that both AoA and word frequency are independent contributors to performance, and should be both taken into account in experimental manipulations. There is also further evidence that subjective familiarity includes other dimensions besides word frequency and may not be appropriate as a substitute for corpora-based frequency counts. Finally, Chapter 4 provides evidence that written, but not spoken frequency, affect visual word recognition. This is important, given that the correlation between spoken and written frequency estimates is not as high as one might assume. This should assuage worries that previous experiments may be compromised since they did not take into account spoken frequency and confirm that, with respect to frequency, only written counts should be taken into account in experiments on visual word recognition.

Theoretically, this thesis provides strong constraints for models of visual word recognition. It shows that models must accommodate both AoA and frequency effects. The AoA findings pose a challenge to standard Parallel Distributed Processing (PDP) accounts of word naming since they predict that AoA effects are eliminated when the input-output mappings are systematic, and this prediction is falsified in the present study. These findings suggest that AoA (and frequency) effects may reflect the structure of lexical-orthographic and lexical-phonological representations themselves. On this latter approach, AoA effects should be observed even under conditions in which input-output mappings are systematic, as is the case of the correspondences between orthography and phonology in quasi-regular languages such as English. This idea is compatible with the growing network model (Steyvers & Tenenbaum, 2005) which proposes that AoA effects stem from a higher degree of centrality of earlier acquired words in the semantic network and can also account for word frequency effects. However, the present results do not necessarily support the idea that semantics is the sole locus of AoA; Experiment 5 found AoA effects in word naming, and it is not clear whether semantics actually plays a role in performance on this task.

The results presented in Chapter 4 indicate that performance on unbalanced-frequency words was similar to performance on the control set matched to their written frequency and significantly different from performance on the control set matched to their spoken frequency. This lends empirical support for models in which the influence of frequency on visual word identification is dependent only on written counts and call into question strong phonological accounts of reading, as the results also imply that access to phonology is not mandatory for tasks that do not explicitly require it, at least for the stages of word recognition affected by word frequency. Those results also imply that the level of interactivity between the orthographic and phonological systems in word recognition is not so profound that the phonological frequency feeds back into visual word identification.

Appendix A.

Appendix A

INSTRUCTIONS FOR THE VARIOUS SUBJECTIVE RATINGS COLLECTED FOR THIS THESIS.

A.1 Instructions for Age of Acquisition Ratings.

Please indicate (in years) the age at which you learned each of the words on the list. An approximate age is good enough for this rating. If you do not know the meaning of a word, just write an X on that space. By "learning a word" we mean the age at which you would have understood that word if somebody had used it in front of you, EVEN IF YOU DID NOT use, read or write it at the time.

If necessary, refer back to these instructions when rating the words. If there are any questions, ask them now. Otherwise, you may begin.

A.2 Instructions for Concreteness Ratings.

(Adapted from Spreen and Schulz, 1966, p.460)

Words may refer to persons, places and things that can be seen, felt, smelled or tasted or to more abstract concepts that cannot be experienced by our senses. The purpose of this experiment is to rate a list of words with respect to concreteness in terms of sense-experience. Any word that refers to objects, materials or persons should receive a high concreteness rating; any word that refers to an abstract concept that cannot be experienced by the senses should receive a low concreteness rating. Think of the words "chair" and "independence". "Chair" can be experienced by our senses and therefore should be rated as high concrete; "independence" cannot be experienced by the senses as such and therefore should be rated as low concrete (or abstract).

You should rate the list of words presented on a 7 point scale, with 7 representing the most concrete extreme and 1 the least. If you do not know a word, or are unsure about it's meaning, please enter an X on that space.

If necessary, refer back to these instructions when rating the words. If there are any questions, ask them now. Otherwise, you may begin.

A.3 Instructions for Imageability Ratings.

(Adapted from Paivio, Yuille, & Madigan, 968, p.4)

Words differ in their capacity to arouse mental images of things and events. Some words arouse a sensory experience, such as a mental picture or sound, very quickly and easily, whereas others may do so only with difficulty (i.e., after a long delay) or not at all. The purpose of this experiment is to rate a list of words as to the ease or difficulty with which they arouse mental images. Any word which, in your estimation, arouses a mental image (i.e., a mental picture, or sound, or other sensory experience) very quickly and easily should be given a high imagery rating; any word that arouses a mental image with difficulty or not at all should be given a low imagery rating. Think of the words "apple" or "fact". Apple would probably arouse an image relatively easily and would be rated as high imagery; fact would probably do so with difficulty and would be rates as low imagery. Since words tend to make you think of other words as associates, e.g., knife-fork, it is important that you note only the ease of getting a mental image of an object or an event to the word.

Your ratings will be made on a seven-point scale, where one is the low imagery end of the scale and seven is the high imagery end of the scale. Make your rating by typing a number from 1 to 7 that best indicates your judgment of the ease or difficulty with which the word arouses imagery. The words that arouse mental images most readily for you should be given a rating of

7; words that arouse images with the greatest difficulty or not at all should be rated 1; words that are intermediate in ease or difficulty of imagery, of course, should be rated appropriately between the two extremes. Fell free to use the entire range of numbers, from 1 to 7; at the same time, don't be concerned about how often you use a particular number as long as it is your true judgment. Work fairly quickly but do not be careless in your ratings.

If necessary, refer back to these instructions when rating the words. If there are any questions, ask them now. Otherwise, you may begin.

A.4 Instructions for Familiarity Ratings.

(Adapted from Gilhooly & Logie, 1980a, p. 396)

This is an experiment to find out how often you have come in contact with certain words. You will be given a list of words and you are to rate each one as to the number of times that you experienced it by simply writing down a number according to a 1 to 7 scale. In this scale, 1 represents "NEVER", that is, you have never seen or heard or used the word in your life; the number 2 represents "RARELY", that is you have seen or heard or used the word at least once before, but only rarely and so on until 7, which represents "VERY OFTEN", that is, you have seen or heard or used the word nearly every day of your life.

Do not be bothered if you are unable to give a definition of some of the words. Simply rate each one as to the number of times you have come in contact with it regardless of its meaning.

There may be some words which you have used or heard more often than you have seen them. Or there may be other words which you have seen more often than you have used or heard them. In such cases, always give the word in the highest rating of the three. For example, you probably use or hear the word "cheers" often, but you may never have seen it in print. In this case, you would rate "cheers" as "Often" and write down the number 6.

When the experimenter tells you to start, go to the list of words and begin rating them at your own speed. This is not a "speed" experiment, each participant will be given plenty of time to finish. On the other hand, do not spend too much time on each word. The important thing is for you to be as accurate as possible.

Be as honest in your ratings as you can. Many of the words in this experiment are very rare, so you are not expected to have come in contact with all of them. Just make the best estimates you are capable of.

If necessary, refer back to these instructions when rating the words. If there are any questions, ask them now. Otherwise, you may begin.

Appendix B.

WORD CHARACTERISTICS AND PERFORMANCE DATA FOR ITEMS IN EXPERIMENT 1:"EFFECT OF WORD FREQUENCY FOR WORDS MATCHED ON AOA".

Table B.1: *Low Frequency Words – LoF Condition.*

	AoA	IMG	FAM	LEN	N	Log WFGF	Log CelexF	Log BNCF	LD _RT	LD _Z	LD_ Acc	NMG _RT	NMG _Z
alchemist	664	394	262	9	0	0.15	0.31	0.18	863	0.44	0.73	748	0.13
baron	472	498	339	5	3	0.85	0.84	0.99	766	-0.16	0.85	616	-0.43
blond	313	580	617	5	4	0.78	1.14	0.77	570	-0.71	0.97	546	-0.77
bump	198	446	523	4	8	1.04	0.85	0.71	615	-0.64	1.00	559	-0.68
idiot	350	423	605	5	1	0.48	1.03	0.86	664	-0.51	1.00	591	-0.70
starve	420	392	571	6	0	0.85	0.88	0.57	802	-0.05	0.85	729	0.09
cereal	231	607	601	6	0	1.00	1.10	0.76	650	-0.47	1.00	759	0.10
confuse	422	260	591	7	0	0.78	0.86	0.79	704	-0.36	1.00	673	-0.27
diner	444	497	442	5	6	0.48	0.44	0.26	632	-0.49	0.97	750	-0.11
sleeve	272	550	560	6	0	0.90	1.02	1.05	624	-0.57	0.91	706	-0.01
exam	459	529	602	4	0	0.48	0.96	0.84	662	-0.43	1.00	609	-0.32
folder	367	563	574	6	6	0.48	0.68	0.62	573	-0.67	0.97	632	-0.56
fork	225	598	584	4	8	1.11	1.17	0.99	650	-0.57	1.00	631	-0.54
glove	228	596	575	5	3	0.70	0.75	0.70	637	-0.52	1.00	598	-0.58
goat	173	637	429	4	7	1.11	1.10	0.89	628	-0.61	0.97	572	-0.61
litter	306	617	558	6	6	0.95	0.95	0.94	726	-0.40	0.82	586	-0.60
junk	403	469	595	4	5	1.08	0.95	0.87	647	-0.52	0.97	626	-0.48
lazy	303	300	587	4	4	1.04	1.13	0.96	531	-0.86	1.00	542	-0.87
lettuce	275	632	587	7	0	0.85	0.87	0.69	632	-0.41	1.00	644	-0.29
olive	458	578	444	5	1	1.00	1.09	1.03	698	-0.48	1.00	591	-0.61
onion	286	617	550	5	1	0.78	1.02	0.88	676	-0.50	1.00	629	-0.56
ounce	501	265	489	5	0	0.70	0.79	0.64	617	-0.57	1.00	625	-0.57
pillow	217	624	602	6	2	1.08	1.17	0.93	601	-0.62	1.00	590	-0.61
plug	242	583	575	4	3	0.78	0.92	0.96	609	-0.63	1.00	569	-0.66

	AoA	IMG	FAM	LEN	N	Log WFGF	Log CelexF	Log BNCF	LD _RT	LD _Z	LD_ Acc	NMG _RT	NMG _Z
potato	233	617	612	6	0	1.18	1.10	0.95	749	-0.22	0.94	657	-0.47
purse	234	637	629	5	5	1.00	1.01	0.91	668	-0.50	0.94	677	-0.27
madness	450	336	564	7	2	0.48	1.14	0.99	700	-0.48	1.00	633	-0.51
rude	274	294	636	4	6	0.85	1.14	0.99	622	-0.52	1.00	590	-0.72
scarf	265	607	602	5	3	0.78	0.95	0.83	696	-0.45	1.00	673	-0.19
sober	675	294	605	5	0	0.60	1.08	0.89	606	-0.49	1.00	613	-0.36
spoon	186	584	612	5	4	1.08	1.14	0.95	575	-0.68	1.00	630	-0.50
starch	489	497	459	6	1	1.00	0.79	0.37	596	-0.56	1.00	675	-0.31
thirst	183	377	584	6	0	0.85	0.84	0.68	629	-0.50	1.00	616	-0.49
vegetable	269	598	591	9	0	1.20	1.38	1.06	775	-0.17	0.97	605	-0.47
wallet	316	602	648	6	4	0.60	0.90	0.85	666	-0.38	1.00	560	-0.62
Mean	337	506	554	5.5	2.7	0.83	0.96	0.81	659	-0.46	0.97	630	-0.44

Note: AoA = Age of Acquisition; IMG = Imageability; FAM = Familiarity; LEN = length in letters; N = orthographic neighbourhood; WFGF = word frequency according to the Educator's Word Frequency Guide; CelexF = word frequency according to CELEX; BNCF = word frequency according to the British National Corpus; LD_RT = mean reaction times for lexical decision; LD_Z = standardized z-scores for lexical decision; LD_Acc = accuracy rate for lexical decision; NMG_RT = mean reaction times for word naming; NMG_Z = standardized z-scores for word naming.

Table B.2: *High Frequency Words – HiF Condition.*

	AoA	IMG	FAM	LEN	N	Log WFGF	Log CelexF	Log BNCF	LD_ RT	LD_ Z	LD_ Acc	NMG _RT	NMG _Z
community	522	416	499	9	0	2.06	2.10	2.39	679	-0.39	1.00	768	0.17
body	267	614	610	4	2	2.64	2.47	2.44	578	-0.63	1.00	584	-0.68
book	214	591	643	4	12	2.46	2.44	2.40	522	-0.89	1.00	601	-0.64
century	414	395	514	7	0	2.12	2.26	2.33	661	-0.48	0.94	674	-0.22
church	278	616	560	6	0	2.01	2.20	2.34	612	-0.74	0.97	602	-0.48
section	480	260	515	7	1	2.01	1.98	2.30	683	-0.37	1.00	648	-0.36
million	419	440	519	7	1	2.17	2.30	2.41	662	-0.40	1.00	632	-0.50
cover	289	443	597	5	10	1.95	2.02	2.06	579	-0.71	0.94	582	-0.61
might	400	346	533	5	8	2.80	2.87	2.78	612	-0.63	0.97	555	-0.70
direct	460	161	516	6	0	1.90	1.99	2.11	587	-0.65	0.94	631	-0.46
theory	557	317	534	6	0	1.94	2.04	2.16	603	-0.62	0.94	626	-0.34
evidence	480	356	504	8	0	1.89	2.18	2.36	589	-0.61	0.97	596	-0.61
produce	431	396	534	7	1	2.20	2.01	2.09	628	-0.55	0.97	641	-0.43
garden	186	635	567	6	2	1.92	2.05	2.06	579	-0.75	1.00	577	-0.61
green	225	609	583	5	3	2.31	2.22	2.18	623	-0.66	1.00	565	-0.71
heart	281	617	578	5	2	2.23	2.16	2.17	573	-0.78	0.97	580	-0.56
market	328	583	518	6	2	1.92	2.13	2.51	594	-0.74	1.00	602	-0.68
village	317	578	524	7	1	2.06	2.13	2.08	663	-0.38	1.00	682	-0.20
appear	357	305	525	6	1	1.94	1.98	2.08	600	-0.67	0.97	619	-0.62
mother	144	638	632	6	1	2.70	2.63	2.44	566	-0.67	1.00	566	-0.69
night	222	607	636	5	8	2.60	2.63	2.55	540	-0.78	0.97	586	-0.74
note	302	512	595	4	9	1.93	1.92	2.05	568	-0.74	1.00	547	-0.74
office	403	613	579	6	0	2.22	2.40	2.43	599	-0.62	0.93	657	-0.47
paper	229	590	635	5	6	2.42	2.24	2.22	623	-0.68	0.97	529	-0.72
park	219	615	582	4	13	1.88	1.84	2.07	616	-0.66	1.00	559	-0.68
picture	219	581	597	7	0	2.39	2.03	2.05	542	-0.69	1.00	591	-0.54
property	447	466	531	8	1	1.89	1.84	2.13	607	-0.59	1.00	624	-0.54
range	436	413	515	5	0	1.89	2.01	2.35	638	-0.60	0.94	596	-0.64
deal	381	383	522	4	16	2.11	2.25	2.16	602	-0.66	1.00	617	-0.48

	AoA	IMG	FAM	LEN	N	Log WFGF	Log CelexF	Log BNCF	LD_ RT	LD_ Z	LD_ Acc	NMG _RT	NMG _Z
road	206	609	604	4	8	2.23	2.33	2.43	630	-0.63	0.97	560	-0.76
capital	380	518	538	7	0	1.95	2.01	2.17	625	-0.55	0.97	622	-0.50
summer	253	618	612	6	2	2.22	2.09	2.08	592	-0.66	0.97	632	-0.42
unit	411	334	513	4	1	1.94	1.81	2.07	621	-0.62	0.94	592	-0.58
water	153	632	641	5	7	3.05	2.64	2.56	565	-0.84	1.00	570	-0.75
woman	258	626	623	5	2	2.27	2.53	2.39	546	-0.73	1.00	581	-0.68
Mean	331	498	564	5.74	3.4	2.18	2.19	2.27	603	-0.64	0.98	606	-0.55

Note: AoA = Age of Acquisition; IMG = Imageability; FAM = Familiarity; LEN = length in letters; N = orthographic neighbourhood; WFGF = word frequency according to the Educator's Word Frequency Guide; CelexF = word frequency according to CELEX; BNCF = word frequency according to the British National Corpus; LD_RT = mean reaction times for lexical decision; LD_Z = standardized z-scores for lexical decision; LD_Acc = accuracy rate for lexical decision; NMG_RT = mean reaction times for word naming; NMG_Z = standardized z-scores for word naming.

Appendix C.

WORD CHARACTERISTICS AND PERFORMANCE DATA FOR ITEMS IN EXPERIMENT 2: "EFFECTS OF AOA FOR WORDS MATCHED ON FREQUENCY".

Table C.1: *Early Acquired Words – EarlyAoA Condition.*

	AoA	IMG	FAM	N	LEN	Log WFGF	Log Celex	Log BNCF	LD_ RT	LD_ Z	LD_ Acc	NMG _RT	NMG _Z
alphabet	244	499	493	0	8	1.20	0.61	0.61	658	-0.43	1.0	633	-0.48
angel	242	554	470	1	5	0.95	1.07	1.24	637	-0.61	1.0	593	-0.63
august	231	386	502	0	6	1.45	1.74	1.94	653	-0.55	1.0	560	-0.67
block	244	483	544	4	5	1.76	1.60	1.67	549	-0.89	1.0	553	-0.65
brave	225	330	545	5	5	1.46	1.31	1.31	573	-0.83	1.0	561	-0.75
bump	198	446	523	8	4	1.04	0.85	0.71	615	-0.64	1.0	559	-0.68
cheese	211	592	588	0	6	1.34	1.46	1.40	579	-0.64	1.0	688	-0.22
chew	210	428	572	4	4	0.90	0.80	0.60	584	-0.52	0.9	662	-0.42
climb	224	485	515	0	5	1.62	1.56	1.39	570	-0.74	0.9	604	-0.44
crawl	216	475	503	3	5	1.04	0.95	0.73	612	-0.57	1.0	631	-0.41
crown	231	644	432	7	5	1.32	1.41	1.76	627	-0.57	1.0	589	-0.56
daisy	219	642	426	2	5	0.85	1.49	0.88	567	-0.83	1.0	593	-0.57
elbow	237	602	564	0	5	1.04	1.22	1.10	632	-0.59	0.9	544	-0.73
elephant	222	616	459	0	8	1.34	1.13	1.01	572	-0.74	1.0	610	-0.57
enjoy	242	362	603	1	5	1.79	1.84	1.83	669	-0.41	0.9	630	-0.44
fairy	242	536	471	4	5	0.90	1.08	0.98	582	-0.71	1.0	677	-0.31
ghost	187	621	466	0	5	1.34	1.31	1.18	651	-0.50	1.0	556	-0.65
goat	173	637	429	7	4	1.11	1.10	0.89	628	-0.61	1.0	572	-0.61
indoors	186	423	578	0	7	0.95	1.06	0.89	598	-0.65	1.0	629	-0.49
juice	250	593	567	1	5	1.30	1.33	1.25	581	-0.77	1.0	655	-0.36
kitten	219	639	517	1	6	1.15	0.70	0.54	611	-0.65	0.9	586	-0.66
nanny	187	524	466	2	5	0.24	0.96	0.79	652	-0.37	1.0	584	-0.52
nasty	225	342	595	4	5	0.70	1.39	1.24	615	-0.67	1.0	595	-0.63
polite	242	257	579	2	6	0.95	1.35	1.13	641	-0.50	1.0	627	-0.61

	AoA	IMG	FAM	N	LEN	Log WFGF	Log Celex	Log BNCF	LD_ RT	LD_ Z	LD_ Acc	NMG _RT	NMG _Z
pray	229	502	456	6	4	0.90	1.21	1.21	624	-0.59	0.9	600	-0.58
rabbit	206	611	523	1	6	1.61	1.07	1.19	590	-0.67	1.0	556	-0.77
shark	242	649	447	6	5	1.04	1.17	0.64	615	-0.57	1.0	617	-0.45
sore	242	358	545	14	4	1.04	1.18	0.95	670	-0.46	1.0	665	-0.32
spin	244	397	487	5	4	1.20	0.96	1.03	524	-0.83	1.0	660	-0.23
thirst	183	377	584	0	6	0.85	0.84	0.68	629	-0.50	1.0	616	-0.49
tickle	209	451	557	5	6	0.30	0.34	0.35	659	-0.45	1.0	644	-0.44
tinsel	229	592	443	0	6	0.15	0.33	0.24	771	0.02	0.8	742	-0.18
tractor	224	651	451	1	7	0.78	0.90	0.78	618	-0.43	0.9	644	-0.44
visit	209	280	567	0	5	1.94	2.00	2.14	569	-0.66	1.0	569	-0.59
zero	223	485	495	1	4	1.30	1.23	1.35	562	-0.76	1.0	589	-0.53
Mean	221	499	513	2.71	5.31	1.11	1.16	1.08	614	-0.60	0.97	611	-0.52

Note: AoA = Age of Acquisition; IMG = Imageability; FAM = Familiarity; LEN = length in letters; N = orthographic neighbourhood; WFGF = word frequency according to the Educator's Word Frequency Guide; CelexF = word frequency according to CELEX; BNCF = word frequency according to the British National Corpus; LD_RT = mean reaction times for lexical decision; LD_Z = standardized z-scores for lexical decision; LD_Acc = accuracy rate for lexical decision; NMG_RT = mean reaction times for word naming; NMG_Z = standardized z-scores for word naming.

Table C.2: *Late Acquired Words – LateAoA Condition.*

	AoA	IMG	FAM	N	LEN	Log WFGF	Log CelexF	Log BNCF	LD_ RT	LD_ Z	LD_ Acc	NMG _RT	NMG _Z
cancer	589	428	510	4	6	1.34	1.88	1.66	617	-0.77	1.0	621	-0.51
canyon	618	566	423	2	6	1.18	0.97	0.58	734	-0.35	0.9	664	-0.28
cigar	504	646	432	0	5	0.78	1.15	0.79	679	-0.46	1.0	674	-0.12
comet	579	524	412	3	5	0.90	0.39	0.56	661	-0.3	0.9	642	-0.32
committee	517	481	532	1	9	1.61	2.07	2.28	722	-0.18	1.0	668	-0.19
corpse	587	535	449	0	6	0.48	1.06	1.00	632	-0.54	1.0	623	-0.44
degree	508	521	574	1	6	1.69	2.02	2.03	679	-0.31	1.0	588	-0.65
drunk	542	451	590	3	5	1.00	1.58	1.38	589	-0.63	1.0	596	-0.64
election	528	435	535	3	8	1.49	1.86	2.02	676	-0.36	1.0	673	-0.44
ensure	624	192	517	3	6	1.00	1.61	2.05	654	-0.4	1.0	705	-0.09
facial	571	269	506	4	4	0.85	1.40	1.82	668	-0.49	0.9	684	-0.31
fund	531	539	461	0	6	0.90	0.80	0.62	709	-0.31	0.9	604	-0.48
gender	700	510	472	0	6	0.60	1.15	1.14	685	-0.39	0.9	637	-0.45
gothic	592	516	520	0	8	0.30	1.14	1.11	590	-0.67	1.0	663	-0.31
graduate	521	294	502	5	5	1.04	1.36	1.43	564	-0.72	1.0	646	-0.41
gross	506	475	521	0	6	1.00	2.03	2.10	650	-0.44	1.0	614	-0.50
income	541	551	500	0	4	1.81	0.98	1.13	600	-0.74	1.0	598	-0.48
jazz	552	596	502	0	6	0.60	0.76	0.85	688	-0.53	1.0	628	-0.47
latter	574	539	684	0	7	1.34	1.47	1.27	553	-0.76	1.0	602	-0.64
lecture	514	559	453	3	4	1.00	0.94	0.89	613	-0.64	1.0	587	-0.70
limb	596	535	486	0	6	0.85	0.92	0.75	632	-0.57	1.0	562	-0.70
liquor	560	621	454	1	7	0.70	1.43	1.01	695	-0.27	1.0	642	-0.48
mistress	505	613	513	1	6	0.90	0.86	0.42	661	-0.45	1.0	632	-0.45
pelvis	550	531	476	4	5	0.30	0.82	0.53	637	-0.53	0.9	683	-0.23
poker	644	483	530	2	6	0.48	1.01	1.14	756	-0.03	0.9	556	-0.71
pride	618	485	507	4	4	1.48	0.90	0.81	669	-0.32	0.9	574	-0.68
sheer	542	285	549	4	4	1.00	1.62	1.52	588	-0.78	1.0	682	-0.30
soul	508	553	466	0	6	1.40	1.18	1.17	626	-0.55	1.0	647	-0.26

	AoA	IMG	FAM	N	LEN	Log WFGF	Log CelexF	Log BNCF	LD_ RT	LD_ Z	LD_ Acc	NMG _RT	NMG _Z
sphere	514	469	500	9	5	1.04	1.18	1.38	650	-0.45	1.0	744	-0.01
stake	580	295	487	6	5	0.90	1.80	1.94	593	-0.59	1.0	653	-0.32
stock	561	512	480	3	5	1.71	0.84	1.15	585	-0.66	1.0	660	-0.30
toxic	572	519	460	3	4	0.70	0.81	0.86	721	-0.25	0.8	599	-0.64
turf	517	516	423	0	5	0.48	0.42	1.07	698	-0.28	0.9	605	-0.55
union	700	473	554	1	4	1.97	0.86	0.72	597	-0.65	0.9	616	-0.43
yoga	571	481	495	1	6	0.19	0.95	0.84	644	-0.49	1.0	605	-0.58
Mean	564	486	499	2.03	5.60	1.00	1.21	1.20	649	-0.48	0.96	634	-0.43

Note: AoA = Age of Acquisition; IMG = Imageability; FAM = Familiarity; LEN = length in letters; N = orthographic neighbourhood; WFGF = word frequency according to the Educator's Word Frequency Guide; CelexF = word frequency according to CELEX; BNCF = word frequency according to the British National Corpus; LD_RT = mean reaction times for lexical decision; LD_Z = standardized z-scores for lexical decision; LD_Acc = accuracy rate for lexical decision; NMG_RT = mean reaction times for word naming; NMG_Z = standardized z-scores for word naming.

Appendix D.

WORD CHARACTERISTICS AND PERFORMANCE DATA FOR ITEMS IN EXPERIMENT 3: "LEXICAL DECISION WITH CHEMISTRY AND PSYCHOLOGY WORDS".

Table D.1: *Psychology Words.*

	PsyF	ChemF	CelexF	WFGF	AoA	Psy_ RT	Psy_ Error	Chem_ RT	Chem_ Error
auditory	279	0	1	4	14.2	584	8.30	706	16.70
bias	310	14	10	4	12.8	599	0.00	615	8.30
cognition	1340	0	0	0	16.5	523	0.00	762	25.00
cue	317	0	6	1	10.7	538	8.30	657	8.30
encoding	300	3	0	0	14.9	515	16.70	640	16.70
explicit	274	4	10	2	13.6	557	0.00	660	8.30
inference	223	0	4	1	15.0	619	8.30	659	8.30
lexical*	419	0	1	0	17.5	(589)	(25.00)	(770)	(66.70)
participant	184	0	3	2	10.9	563	0.00	615	0.00
phonology	168	0	0	0	17.4	564	8.30	726	41.70
priming	558	0	0	1	16.3	585	16.70	668	16.70
rating	126	0	3	7	10.4	609	8.30	596	8.30
retrieval	255	3	1	1	12.0	604	8.30	691	16.70
semantic	515	0	2	1	17.4	587	0.00	795	25.00
serial	285	0	4	1	11.7	543	0.00	620	0.00
stimulus	818	4	11	13	13.2	539	0.00	640	25.00
Mean	398	1.8	3.5	2.4	14.0	569	5.5	670	15.0

Note: PsyF = frequency from the cognitive psychology expert frequency database; ChemF = frequency from the chemistry expert frequency database; CelexF = frequency from the CELEX database (Baayen, et al., 1993); WFGF = frequency from the Educator's Word Frequency Guide (Zeno, et al., 1995); AoA = Age of Acquisition rating (in years); Psy RT = reaction time from psychologists in ms; Psy %Error = error rate for psychologists; Chem RT = reaction time from chemists in ms; Chem %Error = error rate for chemists. All frequencies are given in counts per million.

*This item is included in the means for all lexical variables, but not in the means for RTs and error rates (see details in the Results section of the Experiment).

Table D.2: *Chemistry Words.*

	PsyF	ChemF	CelexF	WFGF	AoA	Psy_ RT	Psy_ Error	Chem_ RT	Chem_ Error
aqueous	0	481	0	1	14.2	850	41.7	588	0.0
carbon	2	565	14	76	11.3	580	0.0	525	0.0
catalyst	0	418	2	2	14	612	0.0	531	0.0
chemistry*	5	1277	13	17	8.5	524	0.0	561	8.3
conformation	0	284	0	0	14	653	0.0	725	0.0
electron	0	937	6	18	12.7	701	16.7	582	8.3
ether	0	327	1	2	14.5	687	58.3	637	0.0
hydrogen	0	1015	13	33	11.3	533	0.0	563	0.0
ion	9	658	2	3	13.6	621	16.7	624	0.0
molecular	2	1030	3	4	13.3	647	0.0	536	8.3
nitrogen	0	372	8	17	11.6	632	8.3	622	0.0
organic	0	895	15	22	12.9	577	8.3	635	0.0
silica	0	227	1	1	15.5	709	41.7	660	8.3
solvent	0	675	2	6	12.7	569	0.0	593	0.0
spectrum	10	544	8	11	11.5	600	0.0	550	0.0
synthesis	8	1165	4	4	14.3	576	0.0	533	0.0
Mean	2.3	680	5.8	13.6	12.9	636	12.8	594	1.7

Note: PsyF = frequency from the cognitive psychology expert frequency database; ChemF = frequency from the chemistry expert frequency database; CelexF = frequency from the CELEX database (Baayen, et al., 1993); WFGF = frequency from the Educator's Word Frequency Guide (Zeno, et al., 1995); AoA = Age of Acquisition rating (in years); Psy RT = reaction time from psychologists in ms; Psy %Error = error rate for psychologists; Chem RT = reaction time from chemists in ms; Chem %Error = error rate for chemists. All frequencies are given in counts per million.

*This item is included in the means for all lexical variables, but not in the means for RTs and error rates (see details in the Results section of the Experiment).

Table D.3: *Early/LoF Words.*

	PsyF	ChemF	CelexF	WFGF	AoA	Psy_ RT	Psy_ Error	Chem_ RT	Chem_ Error
aeroplane	-	-	8	0	4.6	588	0.0	692	0.0
alphabet	-	-	3	15	4.1	528	8.3	605	0.0
banana	-	-	4	5	4.1	531	0.0	546	0.0
bandage	-	-	4	3	5.5	585	16.7	572	8.3
hop	-	-	5	9	4.4	560	16.7	608	8.3
daffodil	-	-	1	0	5.6	610	25.0	698	8.3
dentist	-	-	6	5	4.9	553	0.0	541	8.3
caterpillar	-	-	2	5	4.9	636	0.0	612	0.0
dragon	-	-	8	18	4.8	573	0.0	584	0.0
pony	-	-	8	18	4.8	527	0.0	666	0.0
princess	-	-	12	21	4.1	523	16.7	550	0.0
knitting	-	-	7	5	5.3	602	0.0	574	0.0
shepherd	-	-	6	6	5	545	8.3	618	8.3
spider	-	-	4	17	3.8	496	0.0	536	0.0
strawberry	-	-	3	4	4.4	550	0.0	606	16.7
bubble	-	-	4	6	4.3	578	0.0	572	0.0
Mean			5.3	8.6	4.7	562	5.7	599	3.6

Note: PsyF = frequency from the cognitive psychology expert frequency database; ChemF = frequency from the chemistry expert frequency database; CelexF = frequency from the CELEX database (Baayen, et al., 1993); WFGF = frequency from the Educator's Word Frequency Guide (Zeno, et al., 1995); AoA = Age of Acquisition rating (in years); Psy RT = reaction time from psychologists in ms; Psy %Error = error rate for psychologists; Chem RT = reaction time from chemists in ms; Chem %Error = error rate for chemists. All frequencies are given in counts per million.

Table D.4: *Early/HiF Words.*

	PsyF	ChemF	CelexF	WFGF	AoA	Psy_ RT	Psy_ Error	Chem_ RT	Chem_ Error
adult	-	-	87	47	4.9	509	8.3	549	8.3
afraid	-	-	112	96	5.0	565	0.0	556	0.0
beautiful	-	-	116	148	4.9	474	0.0	569	0.0
beside	-	-	90	87	6.0	555	8.3	574	0.0
daughter	-	-	100	58	4.5	509	0.0	527	0.0
farmer	-	-	31	46	4.4	492	0.0	523	8.3
flower	-	-	28	40	3.5	498	0.0	562	16.7
holiday	-	-	58	14	4.2	465	8.3	543	0.0
kitchen	-	-	106	104	3.8	496	0.0	567	8.3
mountain	-	-	46	118	4.7	506	0.0	615	0.0
quickly	-	-	150	207	4.7	514	0.0	553	0.0
sky	-	-	77	146	3.3	503	0.0	527	0.0
smile	-	-	93	60	3.5	492	0.0	557	0.0
soldier	-	-	26	22	5.3	526	0.0	539	8.3
telephone	-	-	101	72	4.3	528	0.0	555	0.0
television	-	-	114	69	3.4	500	0.0	565	0.0
Mean			83.5	83.4	4.4	508	1.6	555	3.1

Note: PsyF = frequency from the cognitive psychology expert frequency database; ChemF = frequency from the chemistry expert frequency database; CelexF = frequency from the CELEX database (Baayen, et al., 1993); WFGF = frequency from the Educator's Word Frequency Guide (Zeno, et al., 1995); AoA = Age of Acquisition rating (in years); Psy RT = reaction time from psychologists in ms; Psy %Error = error rate for psychologists; Chem RT = reaction time from chemists in ms; Chem %Error = error rate for chemists. All frequencies are given in counts per million.

Table D.5: *List of non-words used in the lexical decision task.*

alency	altor	ancome	barden
cemplain	ambution	aptical	bictional
curriage	essification	decovery	blickade
farkness	cantury	diminate	fantor
gortrayal	castoral	eluction	glatter
inolation	clanting	fex	kad
mirlion	corteinty	finguist	lative
mublic	deg	lote	loute
polonial	mocality	masement	lunding
rinishing	mortion	miction	onactment
risposed	niffusion	peries	plinted
rombination	pertify	pirculation	ponding
sorvant	ractor	refage	procussion
tagnitude	rollector	regument	ralius
tid	sostem	rivergent	rolation
toxt	teliance	tocational	skall

Appendix E.

WORD CHARACTERISTICS AND PERFORMANCE DATA FOR ITEMS IN EXPERIMENT 4: "LEXICAL DECISION WITH GEOLOGY WORDS".

Table E.1: Geology Words

	GeoF	CelexF	WFGF	AoA	Conc	RT	%Error
basalt	316	0	1	10.8	6.1	539	0
carbonate	551	0	2	13.3	5.7	569	0
erosion	452	9	10	11.0	4.8	585	5
fluid	400	14	24	10.1	5.3	556	0
garnet	356	1	1	12.8	6.0	547	0
granite	318	6	8	11.1	6.4	537	5
isotope	677	0	0	13.8	4.8	573	0
magma	291	0	1	12.2	6.3	532	0
mantle	670	4	5	11.3	5.1	542	0
mineral	552	6	26	10.7	6.0	544	5
plateau	299	6	9	10.4	5.5	577	0
quartz	569	1	5	10.3	6.2	524	5
sediment	904	2	4	10.3	5.8	516	5
shear	605	1	0	12.8	4.3	557	5
strata	396	4	2	13.3	5.6	561	10
zircon	512	0	0	14.0	5.9	592	10
Mean	492	3.5	6.2	11.7	5.6	553	3.1

Note: GeoF = frequency from the geology expert frequency database; CelexF = frequency from the CELEX database (Baayen, et al., 1993); WFGF = frequency from Educator's Word Frequency Guide (Zeno, et al., 1995); AoA = Age of Acquisition rating (in years); Conc = concreteness rating; RT = reaction time from geologists in ms; %Error = error rate for geologists. All frequencies are given in counts per million.

Table E.2: *Early/LoF Words.*

	GeoF	CelexF	WFGF	AoA	Conc	RT	%Error
aeroplane	-	8	0	4.4	6.8	622	5
alphabet	-	3	15	4.1	4.9	548	0
balloon	-	3	28	4.3	4.9	549	0
bite	-	17	21	4.3	4.6	601	10
bubble	-	4	6	4.4	5.7	547	15
butterfly	-	5	9	4.2	6.7	563	0
daffodil	-	1	0	4.8	6.7	641	20
dragon	-	8	18	4.4	4.3	536	15
fairy	-	11	7	4.3	4.6	547	0
kitten	-	4	13	3.8	6.6	545	0
pony	-	8	18	4.3	6.2	536	5
princess	-	12	21	4.6	5.5	562	0
puppy	-	5	15	3.7	6.8	566	0
rhyme*	-	2	4	4.8	3.7	596	35
spider	-	4	17	3.8	6.8	508	10
umbrella	-	11	9	4.8	6.7	583	5
Mean		6.6	12.6	4.3	5.7	563.5	5.7

Note: GeoF = frequency from the geology expert frequency database; CelexF = frequency from the CELEX database (Baayen, et al., 1993); WFGF = frequency from the Educator's Word Frequency Guide (Zeno, et al., 1995); AoA = Age of Acquisition rating (in years); Conc = concreteness rating (in a 7 point scale) All frequencies are given in counts per million.
*This item is included in the means for all lexical variables, but not in the means for RTs and error rates (see details in the Results section of the Experiment).

Table E.3: *Early/LoF Words*

	GeoF	CelexF	WFGF	AoA	Conc	RT	%Error
between	-	742	635	5.5	2.8	531	0
book	-	275	290	3.3	6.5	523	0
children	-	656	478	4.7	6.3	512	0
daughter	-	100	58	4.0	5.9	521	5
friend	-	172	173	4.0	4.8	511	10
house	-	559	645	3.7	6.7	496	0
kitchen	-	106	104	3.9	6.6	589	0
morning	-	302	301	3.9	4.8	566	5
party	-	373	158	4.2	5.0	503	5
people	-	1465	2283	4.5	6.4	512	0
picture	-	106	242	4.4	5.8	527	5
school	-	390	579	3.7	5.9	513	0
street	-	254	184	4.6	5.6	559	0
telephone	-	101	72	4.3	6.9	554	0
together	-	366	467	5.3	2.7	567	0
water	-	433	1125	3.5	6.5	502	0
Mean		400	487	4.2	5.6	530	1.9

Note: GeoF = frequency from the geology expert frequency database; CelexF = frequency from the CELEX database (Baayen, et al., 1993); WFGF = frequency from the Educator's Word Frequency Guide (Zeno, et al., 1995); AoA = Age of Acquisition rating (in years); Conc = concreteness rating (in a 7 point scale) All frequencies are given in counts per million.

Table E.4: *List of non-words used in the lexical decision task*

bosin	foll	lind
redge	camb	cluse
luyer	ipple	krother
carlon	glope	mesic
sulnur	curcle	vaice
rebris	puice	accodent
cruter	jutton	ramily
kossil	pullow	mether
parite	pocato	semond
urystal	crother	tumber
yoastal	shiulder	mavourite
placier	ernament	goneral
calcote	tountain	tillion
paldera	enephant	hintory
vectonic	legetable	aserican
evolation	reighbour	quention

Appendix F.

WORD CHARACTERISTICS AND PERFORMANCE DATA FOR ITEMS IN EXPERIMENT 5: "WORD NAMING WITH GEOLOGY WORDS".

Table F.1: *Geology Words.*

	GeoF	CelexF	WFGF	AoA	IMG	RT
basalt	316	0	1	13.8	6.6	477
bearing	262	26	13	10.8	3.9	489
carbonate	551	0	2	14.4	5.7	495
garnet	356	1	1	13.5	6.2	471
glacial	342	1	2	10.4	6.0	470
gradient	163	1	1	10.2	5.4	490
granite	318	6	8	14.0	6.8	491
isotope	677	0	0	15.8	2.7	479
mineral	552	6	26	10.8	5.6	442
plateau	299	6	9	11.8	6.0	483
ridge	348	16	17	10.2	6.1	466
rift	320	2	1	12.3	5.7	459
sample	816	10	30	10.0	3.0	459
sediment	904	2	4	13.0	6.1	477
shear	605	1	0	12.1	5.2	458
strata	396	4	2	14.6	6.3	469
Mean	452	5.2	7.3	12.3	5.4	473

Note: GeoF = frequency from the geology expert frequency database; CelexF = frequency from the CELEX database (Baayen, et al., 1993); WFGF = frequency from Educator's Word Frequency Guide (Zeno, et al., 1995); AoA = Age of Acquisition rating (in years); Img = Imageability rating; RT = reaction time from geologists in ms. All frequencies are given in counts per million.

Table F.2: *Early/LoF Words.*

	GeoF	CelexF	WFGF	AoA	IMG	RT
bite	-	17	21	4.0	5 1	501
bubble	-	4	6	3.7	7 0	501
caterpillar	-	2	5	4.2	6 8	502
glue	-	3	10	4.3	6 0	471
gorilla	-	2	3	4 9	6.8	502
greed	-	8	2	6.0	2.7	466
grumble	-	2	1	5.8	2 5	492
icing	-	2	1	5.5	6.2	467
muddle	-	5	0	5.8	3.1	457
plasticine	-	1	0	4.5	6.2	511
rainbow	-	6	9	4.2	7.0	447
riddle	-	2	3	5.8	3.2	468
shepherd	-	6	6	4.2	6.4	473
somersault	-	1	1	5.4	6.2	483
stripe	-	2	1	4.8	6.3	464
swan	-	5	3	4.8	6.6	443
Mean		4.3	4.5	4.9	6.5	478

Note: GeoF = frequency from the geology expert frequency database; CelexF = frequency from the CELEX database (Baayen, et al., 1993); WFGF = frequency from Educator's Word Frequency Guide (Zeno, et al., 1995); AoA = Age of Acquisition rating (in years). Img = Imageability rating; RT = reaction time from geologists in ms; All frequencies are given in counts per million

Table F.3: *Early/HiF Words.*

	GeoF	CelexF	WFGF	AoA	IMG	RT
because	-	1320	1078	4.6	1.0	469
between	-	742	635	5.5	2.7	464
country	-	367	390	5.2	4.5	467
give	-	484	427	3.6	2.8	469
glass	-	125	129	4.4	6.8	466
great	-	667	800	4.6	2.2	468
group	-	305	361	5.7	3.8	474
important	-	369	610	5.6	1.8	456
mother	-	428	502	3.1	6.5	440
please	-	124	73	3.3	1.7	458
remember	-	256	222	5.3	1.3	471
right	-	826	777	4.2	2.3	464
second	-	340	331	5.5	2.9	447
short	-	193	223	4.6	4.8	455
sound	-	167	241	5.2	2.2	431
street	-	254	184	4.5	6.3	424
Mean		435	436	4.7	3.4	458

Note: GeoF = frequency from the geology expert frequency database; CelexF = frequency from the CELEX database (Baayen, et al., 1993); WFGF = frequency from Educator's Word Frequency Guide (Zeno, et al., 1995); AoA = Age of Acquisition rating (in years); Img = Imageability rating; RT = reaction time from geologists in ms. All frequencies are given in counts per million.

Appendix G.

WORD CHARACTERISTICS AND PERFORMANCE DATA FOR ITEMS IN EXPERIMENT 6: "WORDS WITH HIGH WRITTEN FREQUENCY AND LOW SPOKEN FREQUENCY".

Table G.1: *WrittenHi Words.*

	LogWr	LogSp	Sp_f	Wr_f	AoA	IMG	Len	N	Bf-Tk	Bf-Tp	RT	Error
art	2.23	1.46	29.0	169.3	279	493	3	8	2440	3	486	9.4
aware	1.84	1.16	14.5	68.9	379	298	5	2	1966	30	459	6.3
base	2.07	1.40	24.9	116.8	369	436	4	15	1850	28	492	3.1
below	2.19	1.34	21.9	154.8	257	317	5	0	692	25	489	12.5
Britain	2.43	1.20	15.7	270.7	250	520	7	0	744	55	493	9.4
chapter	2.21	1.06	11.4	162.3	273	340	7	3	1393	124	478	3.1
city	2.38	1.36	22.8	242.4	243	605	4	2	2644	9	469	0.0
danger	1.80	1.00	10.0	63.6	300	505	6	10	2017	136	454	6.3
entry	1.75	0.55	3.6	55.7	285	356	5	0	574	16	469	9.4
event	2.05	0.61	4.0	112.3	264	334	5	1	822	17	474	9.4
fear	1.98	1.05	11.1	96.4	329	394	4	12	2347	32	453	3.1
former	2.27	0.37	2.4	186.7	329	283	6	5	2061	122	499	3.1
further	2.59	1.74	54.4	385.6	354	300	7	1	1750	113	496	3.1
minister	2.42	1.23	17.1	263.3	342	584	8	1	889	100	527	6.3
model	2.15	1.40	24.9	142.6	267	536	5	4	875	35	452	3.1
net*	1.85	1.03	10.7	70.9	269	540	3	16	1906	8	(505)	(21.9)
personal	2.28	1.40	24.9	188.9	293	408	8	1	657	54	494	0.0
popular	2.06	1.17	14.7	115.0	285	422	7	0	293	27	471	3.1
previous	2.12	0.98	9.5	130.9	364	276	8	1	707	65	503	9.4
primary	2.00	1.16	14.5	101.0	297	367	7	1	956	68	494	3.1
prison	1.84	1.00	10.0	68.8	260	593	6	2	845	51	485	9.4
private	2.27	1.52	32.8	185.8	264	432	7	1	770	58	464	3.1
return	2.26	1.24	17.3	180.8	300	293	6	0	1115	47	466	9.4
rise	2.06	1.08	12.1	114.0	292	451	4	13	843	18	509	3.1
royal	2.21	1.17	15.0	160.6	267	480	5	1	587	23	472	6.3
sentence	1.78	0.96	9.0	60.2	250	307	8	0	777	69	485	6.3
smile	1.91	1.18	15.2	81.6	208	615	5	2	1331	27	466	0.0
target	1.86	0.72	5.2	71.9	293	507	6	0	936	82	477	0.0
violence	1.78	0.61	4.0	59.6	293	447	8	0	571	44	531	9.4
Mean	2.09	1.11	16.0	140.7	292	429	5.8	3.5	1219	51	482	5.4

Note: Wr_f = written frequency; Sp_f = spoken frequency; AoA = Age of Acquisition; IMG = Imageability; LEN = length in letters; BF_TK = length-sensitive bigram frequency by tokens; BF_TP = length-sensitive bigram frequency by types; Diff(WrittenHi-HiFreq) = absolute value of the difference between the WrittenHi condition and the HiFreq condition for each variable; Diff(WrittenHi-LoFreq) = absolute value of the difference between the WrittenHi condition and the LoFreq condition for each variable.

*This item is included in the means for all lexical variables, but not in the means for RTs and error rates (see details in the Results section of the Experiment).

Table G.2: *HiFreq Words.*

	LogWr	LogSp	Sp_f	Wr_f	AoA	IMG	Len	N	Bf-Tk	Bf-Tp	RT	Error
actual	1.80	1.75	56.3	63.4	346	259	6	0	731	25	485	9.4
afraid	1.78	1.71	50.8	59.7	448	451	6	0	278	22	488	0.0
break	1.97	2.06	113.7	93.0	223	398	5	6	1280	35	534	9.4
course	2.29	2.26	181.9	193.9	314	391	6	2	1358	51	505	3.1
difference	2.04	2.00	100.4	110.5	378	293	10	0	567	58	509	9.4
easy	2.17	2.12	131.1	148.4	238	321	4	2	1215	16	457	3.1
find	2.61	2.65	446.2	410.1	272	370	4	13	2222	30	481	0.0
floor	2.07	2.15	140.3	117.6	231	544	5	3	834	39	454	3.1
football	1.81	1.90	79.3	65.0	233	597	8	1	159	35	456	6.3
green	2.17	2.21	163.1	149.0	225	609	5	4	1675	42	443	3.1
idea	2.33	2.27	186.4	212.6	292	319	4	2	350	3	469	0.0
insurance	1.85	1.79	61.7	71.1	508	365	9	0	535	69	502	6.3
lady	1.99	2.05	111.6	97.0	231	571	4	4	1070	16	481	6.3
letter	2.13	2.10	127.0	136.2	256	595	6	6	2833	151	457	9.4
lost	2.29	2.23	167.9	196.3	200	300	4	14	3606	31	468	3.1
massive	1.65	1.61	40.4	44.9	273	393	7	2	640	59	485	12.5
matter	2.29	2.36	227.7	197.2	411	298	6	11	3122	160	482	0.0
middle	2.11	2.13	133.5	129.6	193	287	6	6	1187	67	514	3.1
moment	2.33	2.26	183.6	213.8	350	334	6	1	1040	41	461	3.1
month	2.18	2.14	139.6	150.2	243	448	5	1	964	34	474	9.4
must	2.87	2.86	717.1	740.5	209	247	4	15	3657	26	480	3.1
noise	1.67	1.71	51.1	46.7	193	387	5	4	1727	28	453	6.3
notice	1.97	1.91	81.0	92.9	369	467	6	1	916	40	461	9.4
park	2.07	2.09	123.7	117.0	219	573	4	15	1605	36	468	3.1
picture	2.04	1.99	97.4	110.2	219	581	7	0	607	45	465	3.1
square	1.85	1.86	71.7	71.2	250	610	6	1	604	27	460	3.1
touch	1.83	1.87	73.9	68.1	269	456	5	4	3417	41	477	3.1
win	2.03	1.95	90.0	108.4	269	454	3	15	123	7	488	6.3
worth	2.09	2.16	144.9	122.0	369	275	5	2	2676	35	495	12.5
Mean	2.08	2.07	148.0	149.5	284	420	5.6	4.7	1414	44	478	5.2

Note: Wr_f = written frequency; Sp_f = spoken frequency; AoA = Age of Acquisition; IMG = Imageability; LEN = length in letters; BF_TK = length-sensitive bigram frequency by tokens; BF_TP = length-sensitive bigram frequency by types; Diff(WrittenHi-HiFreq) = absolute value of the difference between the WrittenHi condition and the HiFreq condition for each variable; Diff(WrittenHi-LoFreq) = absolute value of the difference between the WrittenHi condition and the LoFreq condition for each variable.

Table G.3: *LoFreq Words.*

	LogWr	LogSp	Sp_f	Wr_f	AoA	IMG	Len	N	Bf-Tk	Bf-Tp	RT	Error
admit	1.57	1.62	42.0	37.5	307	187	5	1	210	10	491	9.4
apricot	0.48	0.58	3.8	3.0	386	591	7	0	272	24	517	6.3
balloon	0.83	0.82	6.7	6.7	233	583	7	0	591	71	486	6.3
blame	1.50	1.55	35.1	31.6	328	356	5	5	784	35	479	3.1
blouse*	0.74	0.70	5.0	5.5	292	595	6	0	1091	37	(528)	(25.0)
catalogue	1.41	1.34	22.1	25.7	300	550	9	0	140	31	518	6.3
cheek	1.30	1.29	19.7	19.9	267	561	5	4	2830	30	476	6.3
cloth	1.30	1.31	20.2	20.1	279	547	5	2	1170	40	516	9.4
comedy	1.18	1.14	13.8	15.0	346	489	6	1	724	39	493	3.1
damp	1.31	1.24	17.6	20.4	264	313	4	10	1028	20	497	9.4
delicious	1.06	1.09	12.4	11.6	221	293	9	1	585	75	474	6.3
desperate	1.43	1.38	24.0	26.9	327	207	9	0	600	95	494	9.4
dive	0.83	0.77	5.9	6.8	322	586	4	12	2696	16	479	6.3
divide	1.21	1.23	17.1	16.4	279	300	6	1	597	26	505	3.1
envelope	1.16	1.24	17.6	14.6	300	554	8	1	235	35	500	0.0
fortune	1.31	1.39	24.7	20.3	375	459	7	0	759	52	475	6.3
kit	1.28	1.22	16.6	18.9	320	367	3	12	320	7	489	15.6
mile	1.51	1.55	35.1	32.4	314	511	4	22	1637	30	498	15.6
olive	0.95	0.96	9.0	9.0	458	578	5	1	423	21	486	3.1
polish	1.32	1.35	22.6	20.8	336	494	6	2	648	49	484	3.1
powder	1.10	1.17	14.7	12.6	300	524	6	2	1892	122	503	0.0
pretend	1.11	1.12	13.3	12.9	214	186	7	1	1083	71	507	9.4
punch	1.19	1.15	14.2	15.6	306	527	5	5	1390	35	503	6.3
quote*	1.09	1.04	10.9	12.3	427	260	5	2	772	21	(563)	(25.0)
slice	1.04	1.10	12.6	11.0	292	507	5	5	1755	34	534	6.3
socket*	0.78	0.74	5.5	6.0	215	550	6	5	1122	77	(550)	(31.3)
spice	0.45	0.55	3.6	2.8	308	592	5	7	1859	32	492	6.3
wheel	1.43	1.40	24.9	26.7	238	576	5	0	4260	31	468	9.4
whistle	0.85	0.82	6.7	7.1	271	574	7	2	422	56	515	12.5
Mean	1.13	1.13	16.5	16.2	304	463	5.9	3.6	1100	42	495	6.9

Note: Wr_f = written frequency; Sp_f = spoken frequency; AoA = Age of Acquisition; IMG = Imageability; LEN = length in letters; BF_TK = length-sensitive bigram frequency by tokens; BF_TP = length-sensitive bigram frequency by types; Diff(WrittenHi-HiFreq) = absolute value of the difference between the WrittenHi condition and the HiFreq condition for each variable; Diff(WrittenHi-LoFreq) = absolute value of the difference between the WrittenHi condition and the LoFreq condition for each variable.

*This item is included in the means for all lexical variables, but not in the means for RTs and error rates (see details in the Results section of the Experiment).

Table G.4: *Nonwords used in the lexical decision task.*

abolity	diminate	lirge	ponding
altor	fature	lote	procussion
ambution	fex	loute	ralius
aptical	fidure	lunding	rask
bictional	flish	mactor	rele
blickade	fot	meldom	resire
boak	frime	melief	ritting
bourt	frisis	miction	rould
breck	gamage	miral	ruture
cafety	ganner	mortion	shirk
cantury	glatter	nane	skall
castoral	glony	neriod	svain
cemplain	goin	niffusion	tid
cirl	gortrayal	nudget	tinal
clanting	guede	onfy	tloth
conrept	irban	panch	toman
corteinty	ixle	pature	toxt
crayer	jense	peison	tragac
culge	kire	peport	wame
curriage	lative	pertify	woral
decovery	leform	plinted	yile
deg	leval	plun	

Appendix H.

WORD CHARACTERISTICS AND PERFORMANCE DATA FOR ITEMS IN EXPERIMENT 7: "WORDS WITH HIGH SPOKEN FREQUENCY AND LOW WRITTEN FREQUENCY".

Table H.1: *SpokenHi Words.*

	LogWr	LogSp	Wr_f	Sp_f	AoA	IMG	Len	N	Bf-Tk	Bf-Tp	RT	Error
bloke	0.81	2.15	6.5	139.9	360	514	5	1	792	33	475	0.0
bloody	1.62	2.89	41.3	771.7	257	514	6	2	320	30	469	6.3
bother	1.27	2.03	18.8	106.1	322	369	6	3	2521	117	450	15.6
chap*	1.04	1.82	10.8	66.5	333	443	4	8	5630	13	(514)	(21.9)
chicken	1.24	2.01	17.5	103.0	250	619	7	1	583	59	440	0.0
crap	0.67	1.85	4.7	71.2	380	243	4	8	138	17	488	3.1
cupboard	1.09	1.80	12.3	63.4	171	633	8	0	181	25	487	0.0
darling	1.35	2.14	22.5	138.0	307	293	7	2	3619	309	513	6.3
eleven	1.39	2.24	24.4	175.7	207	440	6	0	511	46	469	0.0
fridge	0.86	1.78	7.3	60.1	295	620	6	2	467	42	482	3.1
grandma	0.36	1.81	2.3	65.1	147	671	7	1	407	54	526	6.3
hang	1.37	2.22	23.5	166.7	260	537	4	12	3496	35	474	6.3
hate	1.42	2.12	26.3	132.7	278	462	4	16	2661	28	471	6.3
hell	1.66	2.36	45.8	231.3	314	519	4	16	3847	31	479	6.3
hello	1.24	2.60	17.3	394.2	150	327	5	4	513	31	428	0.0
horrible	1.08	2.05	12.0	113.3	171	307	8	1	653	63	471	0.0
lovely	1.64	2.64	43.2	437.9	227	491	6	2	1376	64	461	3.1
minus	0.79	1.86	6.2	72.2	240	293	5	7	593	30	554	15.6
pardon	0.78	2.32	6.0	209.0	342	355	6	1	909	66	524	12.5
pence*	0.83	1.89	6.8	76.9	227	521	5	4	961	30	(452)	(25.0)
pound	1.58	2.69	38.4	494.2	308	553	5	7	3140	37	457	3.1
quid*	0.62	2.33	4.2	212.3	280	547	4	6	1062	8	(590)	(31.3)
reckon	1.11	2.25	13.0	178.3	321	147	6	1	1245	63	544	15.6
silly	1.38	2.09	23.9	124.0	187	307	5	7	781	47	490	6.3
stuff	1.68	2.60	47.8	402.5	240	280	5	4	1008	26	464	9.4
stupid	1.43	2.24	27.0	172.2	200	381	6	0	476	34	456	0.0
tape	1.55	2.48	35.7	304.9	406	573	4	12	640	18	493	6.3
telly	0.65	1.96	4.5	91.9	153	667	5	6	692	47	528	18.8
toilet	1.09	2.19	12.4	156.0	160	603	6	2	755	71	468	3.1
Mean	1.16	2.19	19.4	197.6	258	456	5.5	4.7	1379	51	483	5.9

Note: Wr_f = written frequency; Sp_f = spoken frequency; AoA = Age of Acquisition; IMG = Imageability; LEN = length in letters; BF_TK = length-sensitive bigram frequency by tokens; BF_TP = length-sensitive bigram frequency by types; Diff(WrittenHi-HiFreq) = absolute value of the difference between the WrittenHi condition and the HiFreq condition for each variable; Diff(WrittenHi-LoFreq) = absolute value of the difference between the WrittenHi condition and the LoFreq condition for each variable.
*This item is included in the means for all lexical variables, but not in the means for RTs and error rates (see details in the Results section of the Experiment).

Table H.2: *HiFreq Words*

	LogWr	LogSp	Wr_f	Sp_f	AoA	IMG	Len	N	Bf-Tk	Bf-Tp	RT	Error
before	2.95	2.87	895.1	736.3	208	255	6	0	1897	37	431	3.1
black	2.41	2.38	258.5	237.4	200	589	5	5	1327	44	443	6.3
break	1.97	2.06	93.0	113.7	223	398	5	6	1280	35	438	0.0
course	2.29	2.26	193.9	181.9	314	391	6	2	1358	51	456	3.1
difference	2.04	2.00	110.5	100.4	378	293	10	0	567	58	482	0.0
easy	2.17	2.12	148.4	131.1	238	321	4	2	1215	16	451	6.3
expect	2.02	2.04	104.2	109.9	267	164	6	1	524	21	482	0.0
find	2.61	2.65	410.1	446.2	272	370	4	13	2222	30	493	3.1
floor	2.07	2.15	117.6	140.3	231	544	5	3	834	39	465	6.3
green	2.17	2.21	149.0	163.1	225	609	5	4	1675	42	462	0.0
house	2.71	2.76	513.8	572.3	326	606	5	6	3679	42	473	0.0
idea	2.33	2.27	212.6	186.4	292	319	4	2	350	3	445	0.0
insurance	1.85	1.79	71.1	61.7	508	365	9	0	535	69	506	3.1
lady	1.99	2.05	97.0	111.6	231	571	4	4	1070	16	443	3.1
letter	2.13	2.10	136.2	127.0	256	595	6	6	2833	151	473	3.1
lost	2.29	2.23	196.3	167.9	200	300	4	14	3606	31	474	3.1
matter	2.29	2.36	197.2	227.7	411	298	6	11	3122	160	498	3.1
middle	2.11	2.13	129.6	133.5	193	287	6	6	1187	67	483	9.4
moment	2.33	2.26	213.8	183.6	350	334	6	1	1040	41	464	0.0
month	2.18	2.14	150.2	139.6	243	448	5	1	964	34	460	6.3
mother	2.43	2.37	271.2	233.7	144	638	6	2	2719	117	445	0.0
park	2.07	2.09	117.0	123.7	219	573	4	15	1605	36	463	6.3
picture	2.04	1.99	110.2	97.4	219	581	7	0	607	45	437	6.3
quiet	1.81	1.90	64.6	78.8	223	426	5	2	711	25	476	3.1
rain	1.83	1.78	67.2	60.6	211	618	4	11	1796	28	468	6.3
run	2.34	2.33	220.4	214.0	187	543	3	16	387	14	463	3.1
square	1.85	1.86	71.2	71.7	250	610	6	1	604	27	471	6.3
train	1.91	1.94	81.9	87.3	220	685	5	6	1247	44	457	0.0
white	2.41	2.36	256.0	227.8	214	566	5	3	4195	33	453	6.3
Mean	2.19	2.19	195.1	188.5	257	459	5.4	4.9	1557	47	464	3.3

Note: Wr_f = written frequency; Sp_f = spoken frequency; AoA = Age of Acquisition; IMG = Imageability; LEN = length in letters; BF_TK = length-sensitive bigram frequency by tokens; BF_TP = length-sensitive bigram frequency by types; Diff(WrittenHi-HiFreq) = absolute value of the difference between the WrittenHi condition and the HiFreq condition for each variable; Diff(WrittenHi-LoFreq) = absolute value of the difference between the WrittenHi condition and the LoFreq condition for each variable.

Table H.3: *LoFreq Words.*

	LogWr	LogSp	Wr_f	Sp_f	AoA	IMG	Len	N	Bf-Tk	Bf-Tp	RT	Error
admit	1.57	1.62	37.5	42.0	233	583	5	1	210	10	481	6.3
alarm	1.37	1.35	23.6	22.4	357	406	5	0	1364	40	478	0.0
bake*	0.70	0.72	5.0	5.2	192	562	4	17	2593	22	(531)	(34.4)
balloon	0.83	0.82	6.7	6.7	193	387	7	0	591	71	466	3.1
blame	1.50	1.55	31.6	35.1	215	604	5	5	784	35	467	0.0
blanket	1.06	1.03	11.5	10.7	177	611	7	1	417	64	496	3.1
bunch	1.09	1.08	12.4	12.1	183	599	5	6	1464	38	483	12.5
cinema	1.30	1.22	20.2	16.6	433	229	6	0	281	26	463	0.0
count	1.62	1.64	41.9	43.5	211	582	5	3	3600	47	459	6.3
cousin	1.27	1.29	18.7	19.7	319	525	6	0	1002	47	491	6.3
crawl*	0.64	0.68	4.4	4.7	221	293	5	3	473	38	(512)	(25.0)
crazy	1.26	1.32	18.3	21.1	317	486	5	1	464	37	478	6.3
damp	1.31	1.24	20.4	17.6	246	495	4	10	1028	20	494	15.6
delicious	1.06	1.09	11.6	12.4	220	557	9	1	585	75	474	3.1
dive	0.83	0.77	6.8	5.9	250	587	4	12	2696	16	501	6.3
divide	1.21	1.23	16.4	17.1	233	570	6	1	597	26	495	0.0
dozen*	1.44	1.50	27.4	31.3	261	576	5	3	754	21	(520)	(21.9)
handle	1.59	1.58	39.0	37.8	192	574	6	2	1478	82	462	0.0
luck	1.50	1.57	31.9	37.0	322	586	4	12	1433	22	479	9.4
muddle*	0.27	0.37	1.9	2.4	250	179	6	7	1090	60	(532)	(28.1)
noise	1.67	1.71	46.7	51.1	264	386	5	4	1727	28	468	6.3
pasta	0.76	0.74	5.8	5.5	279	300	5	3	501	34	496	9.4
pension	1.62	1.54	41.4	34.9	292	399	7	1	911	66	506	6.3
pin	1.11	1.17	12.8	14.7	208	488	3	18	111	8	496	6.3
prefer	1.57	1.59	37.2	38.7	307	187	6	0	1673	107	493	9.4
pretend	1.11	1.12	12.9	13.3	448	451	7	1	1083	71	504	0.0
salt	1.48	1.52	29.9	33.0	331	456	4	4	2239	26	513	9.4
taste	1.63	1.69	42.9	48.7	300	327	5	6	939	38	464	9.4
wander	0.83	0.81	6.8	6.4	257	425	6	8	2156	145	495	9.4
Mean	1.21	1.23	21.5	22.3	266	462	5.4	4.5	1181	46	484	5.8

Note: Wr_f = written frequency; Sp_f = spoken frequency; AoA = Age of Acquisition; IMG = Imageability; LEN = length in letters; BF_TK = length-sensitive bigram frequency by tokens; BF_TP = length-sensitive bigram frequency by types; Diff(WrittenHi-HiFreq) = absolute value of the difference between the WrittenHi condition and the HiFreq condition for each variable; Diff(WrittenHi-LoFreq) = absolute value of the difference between the WrittenHi condition and the LoFreq condition for each variable.

*This item is included in the means for all lexical variables, but not in the means for RTs and error rates (see details in the Results section of the Experiment).

Table H.4: *Nonwords used in the lexical decision task.*

alency	farkness	ment	ront
ancome	felly	miction	rould
antial	finker	miral	shalf
barden	flort	mirlion	shirk
boak	forch	mortion	skall
bonge	fragal	mublic	sorvant
bourt	glatter	nane	sostem
breck	gotice	nege	splut
burrel	griss	noken	stimp
cantury	hon	nupper	stiol
cerry	inolation	onfy	tairy
chasel	jense	peries	tase
clenk	jurk	pertify	tont
clond	kire	pounty	toxt
corve	klouse	prace	triut
curriage	lart	prape	tuddle
dalmon	lasp	ractor	wame
darrel	lote	raint	wod
dile	luce	refage	woral
dross	lunding	rillow	yile
fanch	maddle	ritting	zirms
fantor	mapel	rollar	

References

Alario, F. X., & Ferrand, L. (1999). A set of 400 pictures standardized for French: Norms for name agreement, image agreement, familiarity visual complexity, image variability, and age of acquisition. *Behavior Research Methods Instruments & Computers, 31*, 531-552.

Allen, P. A., McNeal, M., & Kvak, D. (1992). Perhaps the lexicon is coded as a function of word frequency. *Journal Of Memory And Language, 31*, 826-844.

Altarriba, J., Bauer, L. M., & Benvenuto, C. (1999). Concreteness, context availability, and imageability ratings and word associations for abstract, concrete, and emotion words. *Behavior Research Methods, Instruments and Computers, 31*, 578-602.

Andrews, S. (1992). Frequency and Neighborhood Effects on Lexical Access: Lexical Similarity or Orthographic Redundancy? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*.

Anisfeld, M. (1964). A comment on "the role of grapheme-phoneme correspondence in the perception of words." *American Journal Of Psychology, 77*, 320-321.

Aston, G., & Burnard, L. (1998). *The BNC Handbook: Exploring the British National Corpus with SARA*. Edimburgh: Edimburgh University Press.

Aveleyra, E., Gomez, C., Ostrosky, F., & Rigalt, C. (1996). Adaptacion de los estimulos no verbales de Snodgrass y Vanderwart en poblacion hispanoparlante: criterios para la denominacion, concordancia de la imagen, familiaridad y complejidad visual. *Revista Mexicana de Psichologia, 13*, 5-19.

Azuma, T. (1996). Familiarity and relatedness of word meanings: Ratings for 110 homographs. *Behavior Research Methods, Instruments and Computers, 28*, 109-124.

Baayen, R. H. (in press). Data mining at the intersection of psychology and linguistics. In A. Cutler (Ed.), *Twenty First Century Psycholinguistics: Four Cornerstones.* Hillsdale, NJ: Erlbaum.

Baayen, R. H., Piepenbrock, R., & Van Rijn, H. (1995). *The CELEX lexical database (CD-ROM).* Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

Balota, D. A. (1990). The role of meaning in word recognition. In D. A. Balota, G. B. F. d'Arcais & K. Rayner (Eds.), *Comprehension processes in reading* (pp. 9-32). Hilldale, N.J.: Erlbaum.

Balota, D. A. (1994). Visual word recognition: The journey from features to meaning. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 303-357). San Diego, CA: Academic Press.

Balota, D. A., Cortese, M. J., Hutchison, K. A., Neely, J. H., Nelson, D., Simpson, G. B., et al. (2002). The English Lexicon Project: A web-based repository of descriptive and behavioral measures for 40,481 English words and nonwords: Washington University.

Balota, D. A., Cortese, M. J., & Pilotti, M. (1999). *Item-level analyses of lexical decision performance: results of a mega-study.* Paper presented at the Abstracts of the 40th annual meeting of the Psychonomics Society, Los Angeles, CA.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual Word Recognition of Single-Syllable Words. *Journal of Experimental Psychology: General, 133*, 283-316.

Balota, D. A., Pilotti, M., & Cortese, M. J. (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition, 29*, 639-647.

Barabasi, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*, 509-512.

Barca, L., Burani, C., & Arduino, L. S. (2002). Word naming times and psycholinguistic norms for Italian nouns. *Behavior Research Methods Instruments & Computers, 34*, 424-434.

Barry, C., Hirsh, K. W., Johnston, R. A., & Williams, C. L. (2001). Age of acquisition, word frequency, and the locus of repetition priming of picture naming. *Journal Of Memory And Language, 44*, 350-375.

Barry, C., Morrison, C. M., & Ellis, A. W. (1997). Naming the Snodgrass and Vanderwart pictures: Effects of age of acquisition, frequency and name agreement. *Quarterly Journal Of Experimental Psychology Section A-Human Experimental Psychology, 50*, 560-585.

Becker, C. A. (1976). Allocation of attention during visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 2*, 556-566.

Becker, C. A. (1980). Semantic context effects in visual word recognition: An analysis of semantic strategies. *Memory & Cognition, 8*, 493-512.

Bell, B. D., Davies, K. G., Hermann, B. P., & Walters, G. (2000). Confrontation naming after anterior temporal lobectomy is related to age of acquisition of the object names. *Neuropsychologia, 38*, 83-92.

Benjafield, J., & Muckenheim, R. (1989). Dates of entry and measures of imagery, concreteness, goodness, and familiarity for 1,046 words sampled from the Oxford English Dictionary. *Behavior Research Methods, Instruments and Computers, 21*, 31-52.

Berman, S., Friedman, D., Hamberger, M., & Snodgrass, J. G. (1989). Developmental picture norms: Relationships between name agreement, familiarity, and visual complexity for child and adult ratings of two sets of line drawings. *Behavior Research Methods, Instruments and Computers, 21*, 371-382.

Berrian, R. W., Metzler, D. P., Kroll, N. E., & Clark-Meyers, G. M. (1979). Estimates of imagery, ease of definition, and animateness for 328 adjectives. *Journal of Experimental Psychology, Human Learning and Memory, 5*, 435-447.

Berry, C. (1971). Advanced frequency information and verbal response times. *Psychonomic Science, 23*, 151-152.

Bird, H., Franklin, S., & Howard, D. (2001). Age of acquisition and imageability ratings for a large set of words, including verbs and function words. *Behavior Research Methods Instruments & Computers, 33*, 73-79.

Bonin, P., Boyer, B., Meot, A., Fayol, M., & Droit, S. (2004). Psycholinguistic norms for action photographs in French and their relationships with spoken and written latencies. *Behavior Research Methods Instruments & Computers, 36*, 127-139.

Bonin, P., Chalard, M., Meot, A., & Fayol, M. (2001). Age-of-acquisition and word frequency in the lexical decision task: Further evidence from the French language. *Cahiers De Psychologie Cognitive-Current Psychology Of Cognition, 20*, 401-443.

Bonin, P., Chalard, M., Meot, A., & Fayol, M. (2002). The determinants of spoken and written picture naming latencies. *British Journal Of Psychology, 93*, 89-114.

Bonin, P., Meot, A., Aubert, L., Malardier, N., Niedenthal, P., & Capelle-Toczek, M. C. (2003). Concreteness, imageability, subjective frequency and emotionality ratings for 866 words. *Annee Psychologique, 103*, 655-694.

Bonin, P., Peereman, R., Malardier, N., Meot, A., & Chalard, M. (2003). A new set of 299 pictures for psycholinguistic studies: French norms for name agreement, image agreement, conceptual familiarity, visual complexity, image variability, age of acquisition, and naming latencies. *Behavior Research Methods Instruments & Computers, 35*, 158-167.

Bowers, J. S., Davis, C. J., & Hanley, D. (in press). Interfering Neighbours: The impact of novel word learning on the identification of visually similar words. *Cognition.*

Bradley, D. C., & Forster, K. I. (1987). A reader's view of listening. *Cognition, 25*, 103-134.

Brown, G. D. A. (1984). A frequency count of 190,000 words in the London-Lund corpus of English conversation. *Behavior Research Methods, Instruments, & Computers, 16*, 502-532.

Brown, G. D. A., & Watson, F. L. (1987). First in, first out: Word Learning Age And Spoken Word-Frequency As Predictors Of Word Familiarity And Word Naming Latency. *Memory & Cognition, 15*, 208-216.

Bruyer, R., & Strypstein, E. (1985). Concreteness and imageability in lateral differences in word perception. *Current Psychology of Cognition, 5*, 111-125.

Brysbaert, M. (1998). Word recognition in bilinguals: Evidence against the existence of two separate lexicons. *Psychologica Belgica, 38*, 163-175.

Brysbaert, M., Lange, M., & Van Wijnendaele, I. (2000). The effects of age-of-acquisition and frequency-of-occurrence in visual word recognition: Further evidence from the Dutch language. *European Journal Of Cognitive Psychology, 12*, 65-85.

Brysbaert, M., Van Wijnendaele, I., & De Deyne, S. (2000). Age-of-acquisition effects in semantic processing tasks. *Acta Psychologica, 104*, 215-226.

Burgess, C., & Livesay, K. (1998). The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kucera and Francis. *Behavior Research Methods Instruments & Computers, 30*, 272-277.

Callejas, A., Correa, A., Lupianez, J., & Tudela, P. Intracategorical Associative Norms For 612 Words From Six Semantic Categories In Spanish / Normas Asociativas Intracategoriales para 612 Palabras de Seis Categorias Semanticas en Espanol. *Psicologica, 24*, 185-200.

Carr, T. H., & Pollatsek, A. (1985). Recognizing printed words: A look at current models. In D. Besner, T. G. Waller & G. E. MacKinnon (Eds.), *Reading research: Advances in theory and practice 5* (pp. 1-82). San Diego, CA: Academic Press.

Carroll, J. B. (1967). On sampling from a lognormal model of word-frequency distribution. In H. Kucera & W. N. Francis (Eds.), *Computational analysis of present day American English* (pp. 406-424). Providence, RI: Brown University Press.

Carroll, J. B. (1970). An alternative to Julliand's usage coefficient for lexical frequencies an a proposal for a Standard Frequency Index (SFI). *Computer Studies in the Humanities and Verbal Behavior, 3*, 61-65.

Carroll, J. B., Davies, P., & Richman, B. (1971). *The American Heritage word frequency book.*New York: American Heritage.

Carroll, J. B., & White, M. N. (1973a). Age of acquisition norms for 220 picturable nouns. *Journal of Verbal Learning and Verbal Behavior, 12*, 563-576.

Carroll, J. B., & White, M. N. (1973b). Word frequency and age-of-acquisition as determiners of picture-naming latency. *Quarterly Journal Of Experimental Psychology, 25*, 85-95.

Chalard, M., Bonin, P., Meot, A., Boyer, B., & Fayol, M. (2003). Objective age-of-acquisition (AoA) norms for a set of 230 object names in French: Relationships with psycholinguistic variables, the English data from Morrison et al. (1997), and naming latencies. *European Journal Of Cognitive Psychology, 15*, 209-245.

Chen, Y., Yen, Y., Tsai, M., & Yeh, Y. (2001). Norms for 105 object contours. *Chinese Journal of Psychology, 43*, 103-113.

Chiarello, C., Shears, C., & Lund, K. (1999). Imageability and distributional typicality measures of nouns and verbs in contemporary English. *Behavior Research Methods, Instruments and Computers, 31*, 603-637.

Clear, J. (1992). Corpus Sampling. In G. Leitner (Ed.), *New directions in English language corpora* (pp. 21-31). Berlin: Mouton de Gruyter.

Colombo, L., & Burani, C. (2002). The influence of age of acquisition, root frequency, and context availability in processing nouns and verbs. *Brain And Language, 81*, 398-411.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology, 33A*, 497-505.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed processing approaches. *Psychological Review, 100*, 589-608.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance* (VI ed., pp. 535-555). New York: Academic Press.

Coltheart, M., Patterson, K., & Marshall, J. C. (Eds.). (1987). *Deep dyslexia* (2nd ed.). London: Rutledge & Kegan Paul.

Coltheart, M., & Rastle, K. (1994). Serial processing and reading aloud: Evidence for dual-route models of reading. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 1197-1211.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. C. (2001). DRC: A Dual Route Cascaded Model of Visual Word Recognition and Reading Aloud. *Psychological Review, 108*, 204-256.

Cortese, M., & Fugett, A. (2004). Imageability Ratings for 3,000 Monosyllabic Words. *Behavior Research Methods, Instruments, & Computers, 36*, 384–387.

Costa, A., Miozzo, M., & Caramazza, A. (1999). Lexical selection in bilinguals: Do words in the bilingual's two lexicons compete for selection? *Journal Of Memory And Language, 41*, 381-391.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Cuetos, F., Aguado, G., Izura, C., & Ellis, A. W. (2002). Aphasic naming in Spanish: predictors and errors. *Brain And Language, 82*, 344-365.

Cuetos, F., & Alija, M. (2003). Normative data and naming times for action pictures. *Behavior Research Methods Instruments & Computers, 35*, 168-177.

Cuetos, F., Ellis, A. W., & Alvarez, B. (1999). Naming times for the Snodgrass and Vanderwart pictures in Spanish. *Behavior Research Methods Instruments & Computers, 31*, 650-658.

Cycowicz, Y. M., Friedman, D., Rothstein, M., & Snodgrass, J. G. (1997). Picture naming by young children: Norms for name agreement, familiarity, and visual complexity. *Journal of Experimental Child Psychology, 65*, 171-237.

Dahl, H. (1979). *Word frequencies of spoken American English.*Essex, CT: Verbatim.

Dale, E. (1948). Dale list of 3,000 familiar words. *Educational Research Bulletin, 27*, 45-54.

Dale, E., & Eichholz, G. (undated). *Children's knowledge of words.*Columbus, Ohio: Bureau of Educational Research and Service, Ohio State University.

Dale, P. S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods Instruments & Computers, 28*, 125-127.

Davis, C. J. (1999). *The Self-Organising Lexical Acquisition and Recognition (SOLAR) model of visual word recognition.* University of New South Wales, Sydney.

Davis, C. J. (2005). N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research, Methods, Instruments, & Computers.*

Davis, C. J., & Perea, M. (in press). BuscaPalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish. *Behavior Research Methods, Instruments, and Computers.*

de Bot, K. (1992). A bilingual production model: Levelt's speaking model adapted. *Applied Linguistics, 13,* 1-24.

De Houwer, J. (1998). The semantic Simon effect. *Quarterly Journal Of Experimental Psychology: Human Experimental Psychology, 51A,* 683-688.

De Moor, W., Ghyselinck, M., & Brysbaert, M. (2000). A validation study of the age-of-acquisition norms collected by Ghyselinck, De Moor, & Brysbaert. *Psychologica Belgica, 40,* 99-114.

Dell'acqua, R., Lotto, L., & Job, R. (2000). Naming times and standardized norms for the Italian PD/DPSS set of 266 pictures: Direct comparisons with American, English, French, and Spanish published databases. *Behavior Research Methods Instruments & Computers, 32,* 588-615.

Desrochers, A., & Bergeron, M. (2000). Value of subjective frequency and imagery for a sample of 1,916 French language nouns / Valeurs de frequence subjective et d'imagerie pour un echantillon de 1,916 substantifs de la langue francaise. *Canadian Journal of Experimental Psychology, 54,* 274-325.

Di Vesta, F. J., & Walls, R. T. (1970). Factor analysis of the semantic attributes of 487 words and some relationships to the conceptual behavior of fifth-grade children. *Journal of Educational Psychology, 61,* 1-15.

Duyck, W., Desmet, T., Verbeke, L., & Brysbaert, M. (2004). WordGen: A tool for word selection and non-word generation in Dutch, German, English and French. *Behavior Research Methods, Instruments and Computers, 36,* 488-499.

Ebel, R. L. (1951). Evaluation of the reliability of ratings. *Psychometrika, 16,* 407-424.

Ellis, A. W., & Lambon-Ralph, M. A. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks. *Journal Of Experimental Psychology: Learning Memory And Cognition, 26,* 1103-1123.

Ellis, A. W., & Morrison, C. M. (1998). Real age-of-acquisition effects in lexical retrieval. *Journal Of Experimental Psychology-Learning Memory And Cognition, 24,* 515-523.

Fear, W. J. (1997). Ratings for Welsh words and their English equivalents. *Behavior Research Methods Instruments & Computers, 29,* 425-445.

Fellbaum, C. (1998). A semantic network of English: The mother of all WordNets. *Computers And The Humanities, 32,* 209-220.

Ferrand, L., Grainger, J., & New, B. (2003). Age-of-acquisition norms for 400 monosyllabic french words. *Annee Psychologique, 103,* 445-467.

Forster, J. C., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior, 12,* 627-635.

Forster, J. C., & Shen, D. (1996). No enemies in the neighborhood: Absence of inhibitory neighborhood effects in lexical decision and semantic categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 696-713.

Forster, K. I. (1976). Accessing the mental lexicon. In R. J. Wales & E. Walker (Eds.), *New approaches to language mechanisms* (pp. 257-287). Amsterdam: North-Holland.

Forster, K. I. (1979). Levels of processing and the structure of the language processor. In W. E. Cooper & E. Walker (Eds.), *Sentence processing* (pp. 27-85). Hillsdale, NJ: Erlbaum.

Forster, K. I. (1989). Basic issues in lexical processing. In W. Marslen-Wilson (Ed.), *Lexical representations and process* (pp. 75-107). Cambridge, MA: MIT Press.

Forster, K. I. (1990). Lexical processing. In D. N. Osherson & H. Lasnick (Eds.), *Language: an invitation to congnitive science* (Vol. 1, pp. 95-131). Cambridge, MA: MIT Press.

Forster, K. I. (1992). Memory-addressing mechanisms and lexical access. In L. Katz & R. Frost (Eds.), *Orthography, phonology, morphology and meaning* (pp. 413-434). North-Holland: Elsevier.

Forster, K. I., & Bednall, E. S. (1976). Terminating an exhaustive search in lexical access. *Memory & Cognition, 4*, 53-61.

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods Instruments & Computers, 35*, 116-124.

Frederiksen, J. R., & Kroll, J. F. (1976). Spelling and sound: Approaches to the internal lexicon. *Journal of Experimental Child Psychology: Human Perception and Performance, 2*, 361-379.

French, N., Carter, C. W., & Koenig, W. (1930). The words and sounds of telephone conversations. *Bell Systems Technical Journal, 9*, 290-324.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences, 3*, 128-135.

Friendly, M., Franklin, P. E., Hoffman, D., & Rubin, D. C. (1982). The Toronto Word Pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods and Instrumentation, 14*, 375-399.

Fromkin, V., & Rodman, R. (1993). *An Introduction to Language.*Fort Worth: Harcourt Brace Jovanovich.

Frost, R. (1998). Toward a Strong Phonological Theory of Visual Word Recognition: True Issues and False Trails. *Psychological Bulletin, 123*, 71-99.

Gardner, M. K., Rothkopf, E. Z., Lapan, R., & Lafferty, T. (1987). The Word-Frequency Effect In Lexical Decision - Finding A Frequency-Based Component. *Memory & Cognition, 15*, 24-28.

Gerhand, S., & Barry, C. (1998). Word frequency effects in oral reading are not merely age-of- acquisition effects in disguise. *Journal Of Experimental Psychology-Learning Memory And Cognition, 24*, 267-283.

Gerhand, S., & Barry, C. (1999a). Age-of-acquisition and frequency effects in speeded word naming. *Cognition, 73*, B27-B36.

Gerhand, S., & Barry, C. (1999b). Age of acquisition, word frequency, and the role of phonology in the lexical decision task. *Memory & Cognition, 27*, 592-602.

Gernsbacher, M. A. (1984). Resolving 20 years of inconsisten interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General, 113*, 256-281.

Ghyselinck, M., Custers, R., & Brysbaert, M. (2003). Age-of-acquisition ratings for 2332 Dutch words from 49 different semantic categories. *Psychologica Belgica, 43*, 181-214.

Ghyselinck, M., Custers, R., & Brysbaert, M. (2004). The effect of age of acquisition in visual word processing: Further evidence for the semantic hypothesis. *Journal Of Experimental Psychology-Learning Memory And Cognition, 30*, 550-554.

Ghyselinck, M., De Moor, W., & Brysbaert, M. (2000). Age-of-acquisition ratings for 2816 Dutch four- and five-letter nouns. *Psychologica Belgica, 40*, 77-98.

Ghyselinck, M., Lewis, M. B., & Brysbaert, M. (2004). Age of acquisition and the cumulative-frequency hypothesis: A review of the literature and a new multi-task investigation. *Acta Psychologica, 115*, 43-67.

Gilhooly, K. J. (1984). Word age-of-acquisition and residence time in lexical memory as factors in word naming. *Current Psychological Research and Reviews, 3*, 24-31.

Gilhooly, K. J., & Gilhooly, M. L. (1979). Age-Of-Acquisition Effects In Lexical And Episodic Memory Tasks. *Memory & Cognition, 7*, 214-223.

Gilhooly, K. J., & Gilhooly, M. L. M. (1980). Validity Of Age-Of-Acquisition Ratings. *British Journal Of Psychology, 71*, 105-110.

Gilhooly, K. J., & Hay, D. (1977). Imagery, Concreteness, Age Of Acquisition, Familiarity, And Meaningfulness Values For 205 5-Letter Words Having Single- Solution Anagrams. *Behavior Research Methods & Instrumentation, 9*, 12-17.

Gilhooly, K. J., & Logie, R. H. (1980a). Age-Of-Acquisition, Imagery, Concreteness, Familiarity, And Ambiguity Measures For 1,944 Words. *Behavior Research Methods & Instrumentation, 12*, 395-427.

Gilhooly, K. J., & Logie, R. H. (1980b). Meaning-Dependent Ratings Of Imagery, Age Of Acquisition, Familiarity, And Concreteness For 387 Ambiguous Words. *Behavior Research Methods & Instrumentation, 12*, 428-450.

Gilhooly, K. J., & Logie, R. H. (1981). Word age-of-acquisition and visual recognition thresholds. *Current Psychological Research, 1*, 215-226.

Gilhooly, K. J., & Watson, F. L. (1981). Word age-of-acquisition effects: A review. *Current-Psychological-Reviews, 1*, 269-286.

Glushko, R. J. (1979). The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance, 5*, 674-691.

Gordon, B. (1985). Subjective frequency and the lexical decision latency function: Implications for mechanisms of lexical access. *Journal of Memory and Language, 24*, 631-645.

Grainger, J. (1990). Word frequency and the neighborhood frequency effects in lexical decision and word naming. *Journal Of Memory And Language, 29*, 228-244.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding: parallel development and coding of neural feature detectors. *Biological Cybernetics, 23*, 121-134.

Guan, Y., & Fang, F. (2002). The effect of word learning age in primary school children's visual Chinese character recognition: II. *Acta Psychologica Sinica, 34*, 23-28.

Hager, W., & Asmuss-Kumke, M. (1996). Norms for degrees of threat and familiarity of German words / Normen der Bedrohlichkeit und der Bekanntkeit von Woertern. *Zeitschrift fuer Klinische Psychologie: Forschung und Praxis, 25*, 324-327.

Harley, T. (2001). *The Psychology of Language: From data to theory.*Hove, U.K.: Psychology Press.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review, 106*, 491-528.

Havens, L. L., & Foote, W. E. (1963). The effect of competition on visual duration threshold and its independence of stimulus frequency. *Journal Of Experimental Psychology, 65*, 6-11.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review, 7*, 185-207.

Hermans, D., & De Houwer, J. (1994). Affective and subjective familiarity ratings of 740 Dutch words. *Psychologica Belgica, 34*, 115-139.

Hino, Y., Lupker, S. J., Ogawa, T., & Sears, C. R. (2003). Masked repetition priming and word frequency effects across different types of Japanese scripts: An examination of the lexical activation account. *Journal of Memory and Language, 48*, 33-66.

Hirsh, K. W., & Funnell, E. (1995). Those old, familiar things: Age of acquisition, familiarity and lexical access in progressive aphasia. *Journal Of Neurolinguistics, 9*, 23-32.

Horn, E. (1926). *A basic writing vocabulary: 10,000 frequently used words in writing.* Iowa City: University of Iowa.

Howes, D. H. (1966). A word count of spoken English. *Journal of Verbal Learning and Verbal Behavior, 5*, 572-606.

Howes, D. H., & Solomon, R. L. (1951). Visual duration threshold as a function of word-probability. *Journal Of Experimental Psychology, 41*, 401-410.

Hulme, C., Stuart, G., Brown, G. D. A., & Morin, C. (2003). High- and low-frequency words are recalled equally well in alternating lists: Evidence for associative effects in serial recall. *Journal Of Memory And Language, 49*, 500-518.

Inhoff, A. W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & Psychophysics, 40*, 431-439.

Izura, C., & Ellis, A. W. (2002). Age of Acquisition effects in word recognition and production in first and second languages. *Psicologica, 23*, 245-281.

Izura, C., & Ellis, A. W. (2004). Age of acquisition effects in translation judgement tasks. *Journal Of Memory And Language, 50*, 165-181.

Jacobs, A. M., & Grainger, J. (1994). Models of Visual Word Recognition - Sampling the State of the Art. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 1311-1334.

Jorm, A. F. (1991). The Validity Of Word Age-Of-Acquisition Ratings - A Longitudinal-Study Of A Childs Word Knowledge. *British Journal Of Developmental Psychology, 9*, 437-444.

Juhasz, B. J., & Rayner, K. (2003). Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading. *Journal Of Experimental Psychology-Learning Memory And Cognition, 29*, 1312-1318.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*, 329-354.

Kawamoto, A. H., Ferrar, W. T., & Kello, C. T. (1994). When two meanings are better than one: Modeling the ambiguity advantage using a recurrent distributed network. *Journal of Experimental Psychology: Human Perception and Performance, 20.*

Kello, C. T., & Kawamoto, A. H. (1998). Runword: An IBM-PC software package for the collection and acoustic analysis of speeded naming responses. *Behavior Research Methods Instruments & Computers, 30*, 371-383.

Kerr, N. H., & Johnson, T. H. (1991). Word norms for blind and sighted subjects: Familiarity, concreteness, meaningfulness, imageability, imagery modality, and word associations. *Behavior Research Methods, Instruments and Computers, 23*, 461-485.

Kessler, B., Treiman, R., & Mullennix, J. (2002). Phonetic biases in voice key response time measurements. *Journal of Memory and Language, 47*, 145-171.

Kohnstamm, G. A., Schaerlaekens, A. M., de Vries, A. K., Akkerhuis, G. W., & Froonincksx, M. (1981). *Nieuwe streeflijst woordenschat.* Lisse, The Netherlands: Swets & Zeitlinger.

Kroll, J. F., & Merves, J. S. (1986). Lexical Access For Concrete And Abstract Words. *Journal Of Experimental Psychology-Learning Memory And Cognition, 12*, 92-107.

Kucera, H., & Francis, W. (1967). *Computational analysis of present day American English*. Providence, RI: Brown University Press.

Kutas, M., & Van Petten, C. (1994). Psycholinguistics electrified: Event-related brain potential investigations. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 83-143). San Diego: Academic Press.

Lambon-Ralph, M. A., Graham, K. S., Ellis, A. W., & Hodges, J. R. (1998). Naming in semantic dementia - What matters? *Neuropsychologia, 36*, 775-784.

Lee, C. J. (2003). Evidence-Based Selection of Word Frequency Lists. *Journal of Speech-Language Pathology and Audiology, 27*, 170-173.

Leech, G. (1993). 100 million words of English: a description of the background, nature and prospects of the British National Corpus. *English Today, 9*.

Lete, B., Sprenger-Charolles, L., & Cole, P. (2004). MANULEX: A grade-level lexical database from French elementary school readers. *Behavior Research Methods Instruments & Computers, 36*, 156-166.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral And Brain Sciences, 22*, 1-75.

Lewis, M. B. (1999a). Age of acquisition in face categorisation: is there an instance-based account? *Cognition, 71*, B23-B39.

Lewis, M. B. (1999b). Are age-of-acquisition effects cumulative-frequency effects in disguise? A reply to Moore, Valentine and Turner (1999). *Cognition, 72*, 311-316.

Lewis, M. B., Gerhand, S., & Ellis, H. D. (2001). Re-evaluating age-of-acquisition effects: are they simply cumulative-frequency effects? *Cognition, 78*, 189-205.

Lipinski, J., & Gupta, P. (2005). Does neighborhood density influence repetition latency for nonwords? Separating the effects of density and duration. *Journal of Memory and Language, 52*, 171-192.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods Instruments & Computers, 28*, 203-208.

Marcel, A. J., & Patterson, K. (1978). Word recognition and production: Reciprocity in clinical and normal studies. In J. Requin (Ed.), *Attention and performance: VII*. Hillsdale, NJ: Erlbaum.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19*, 313-330.

Masterson, J., & Druks, J. (1998). Description of a set of 164 nouns and 102 verbs matched for printed word frequency, familiarity and age-of-acquisition. *Journal Of Neurolinguistics, 11*, 331-354.

McClelland, J. L. (1987). The case for interactionism in language processing. In M. Coltheart (Ed.), *Attention and Performance, XII* (pp. 3-36). London: Lawrence Erlbaum Associates Limited.

McClelland, J. L., & Rumelhart, D. E. (1981). An Interactive Activation Model Of Context Effects In Letter Perception.1. An Account Of Basic Findings. *Psychological Review, 88*, 375-407.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed Memory And The Representation Of General And Specific Information. *Journal Of Experimental Psychology-General, 114*, 159-188.

McCloskey, M., & Cohen, N. (1989). Catastrophic interference in connectionist networks: the sequential learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 24, pp. 109-164). New York: Academic Press.

McCusker, L. M. (1977). Some determinants of word recognition: Frequency, *24th Annual Convention of the Southwestern Psychological Association*. Fort Worth, TX.

McGinnies, E., Comer, P. B., & Lacey, O. L. (1952). Visual recognition thresholds as a function of word lenght and word frequency. *Journal Of Experimental Psychology, 44*, 65-69.

Meng, X., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin, 111*, 172-175.

Metsala, J. L., & Walley, A. C. (1998). Spoken vocabulary growth and the segmental restructuring of lexical representations: Precursors to phonemic awareness and early reading ability. In J. l. Metsala & L. C. Ehri (Eds.), *Word recognition in beginning literacy* (pp. 89-120). Mahwah, New Jersey: Lawrence Erlbaum.

Milgram, S. (1967). The small-world problem. *Psychology Today, 1*, 60-67.

Miller, G. A. (1995). WordNet: An on-line lexical database. *Special Issue of International Journal of Lexicography, 3*.

Monaghan, J., & Ellis, A. W. (2002). Age of acquisition and the completeness of phonological representations. *Reading and Writing: An Interdisciplinary Journal, 15*, 759-788.

Monaghan, J., & Ellis, A. W. (2002). What exactly interacts with spelling-sound consistency in word naming? *Journal Of Experimental Psychology-Learning Memory And Cognition, 28*, 183-206.

Monsell, S. (1991). The nature and locus of word frequency effects in reading. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading* (pp. 148-152). Hillsdale, N.J.: L. Erlbaum Associates.

Monsell, S., Doyle, M. C., & Haggard, P. N. (1989). Effects of frequency on visual word recognition tasks: Where are they? *Journal of Experimental Psychology: General, 118*, 43-71.

Moore, V., & Valentine, T. (1998). The effect of age of acquisition on speed and accuracy of naming famous faces. *Quarterly Journal Of Experimental Psychology Section A-Human Experimental Psychology, 51*, 485-513.

Morris, P. E., & Reid, R. L. (1972). Canadian and British ratings of the imagery values of words. *British Journal of Psychology, 63*, 163-164.

Morrison, C. M. (2003). Interpret with caution: Multicollinearity in multiple regression of cognitive data. *Perceptual And Motor Skills, 97*, 80-82.

Morrison, C. M., Chappell, T. D., & Ellis, A. W. (1997). Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. *Quarterly Journal Of Experimental Psychology Section A-Human Experimental Psychology, 50*, 528-559.

Morrison, C. M., & Ellis, A. W. (1995). Roles Of Word-Frequency And Age Of Acquisition In Word Naming And Lexical Decision. *Journal Of Experimental Psychology-Learning Memory And Cognition, 21*, 116-133.

Morrison, C. M., & Ellis, A. W. (2000). Real age of acquisition effects in word naming and lexical decision. *British Journal Of Psychology, 91*, 167-180.

Morrison, C. M., Ellis, A. W., & Quinlan, P. T. (1992). Age Of Acquisition, Not Word-Frequency, Affects Object Naming, Not Object Recognition. *Memory & Cognition, 20*, 705-714.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review, 76*, 165-178.

Morton, J. (1979). Facilitation in word recognition: Experiments causing change in the logogen model. In P. A. Kolers, M. E. Wrolstad & H. Bouma (Eds.), *Processing models of visible language* (pp. 259-268). New York: Plenum.

Morton, J. (1980). The logogen model and orthographic structure. In U. Frith (Ed.), *Cognitive processes in spelling* (pp. 117-133). London: Academic Press.

Murray, W. S., & Forster, K. I. (2004). Serial mechanisms in lexical access: The rank hypothesis. *Psychological Review, 111*, 721-756.

Nacher, M. J., Gotor, A., & Algarabel, S. (1998). Traducciones equivalentes en cataln y castellano de 1533 palabras, y sus valores normativos en concresion, familiaridad y significatividad. *Psicologica, 19*, 1-26.

Nam, K., Lee, Y., & Lee, C. H. (2004). The locus of word length and frequency effect in comprehending English words by Korean-English bilinguals and Americans. *Lecture Notes in Computer Science, 3316*, 306-315.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods Instruments & Computers, 36*, 402-407.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1-55). Hillsdale, N.J.: Erlbaum.

Nisi, M., Longoni, A. M., & Snodgrass, J. G. (2000). Italian measurement on the relation of name, familiarity, and acquisition age for the 260 figures of Snodgrass and Vanderwart / Misure italiane per l'accordo sul nome, familiarita ed eta di acquisizione per le 260 figure di Snodgrass e Vanderwart (1980). *Giornale Italiano di Psicologia, 27*, 205-218.

Norris, D. (1986). Word recognition: Context effects without priming. *Cognition, 22*, 93-156.

Novick, L. R., & Sherman, S. J. (2004). Type-based bigram frequencies for five-letter words. *Behavior Research Methods, Instruments and Computers, 36*, 397-401.

Offe, H., Anneken, G., & Kessler, E. (1981). Norms of the concreteness and imagery ratings of 234 nouns / Normen fuer die Konkretheits- und Vorstellbarkeitseinschaetzung von 234 Substantiven. *Psychologische Beitrage, 23*, 65-85.

Paap, K. R., Newsome, S. L., McDonald, J. E., & Schvaneveldt, R. W. (1982). An activation-verification model for letter and word recognition: The word superiority effect. *Psychological Review, 89*, 573-594.

Paap, K. R., McDonald, J.E., Schvaneveldt, R. W., & Noel, R. W. (1987). Frequency and pronounceability in visually presented naming and lexical decision tasks. In M. Coltheart (Ed.), *Attention and performance, XII* (pp. 221-243). London: Lawrence Erlbaum Associates Limited.

Paivio, A. (1971). *Imagery and verbal processes.* New York: Holt, Rinehart, & Winston.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness Imagery And Meaningfulness Values For 925 Nouns. *Journal Of Experimental Psychology, 76*, 1-&.

Patterson, K., Marshall, J. C., & Coltheart, M. (Eds.). (1985). *Surface dyslexia: Neuropsychological and cognitive studies of phonological reading.* Hove, UK: Lawrence Erlbaum Associates Ltd.

Patterson, K., Seidenberg, M. S., & McClelland, J. L. (1989). Connections and disconnections: Acquired dyslexia in a computational model of reading processes. In R. G. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neurobiology* (pp. 131-181). Oxford, England: Oxford University Press.

Perea, M. (1993). Una base de palabras de cuatro letras: Indices de frecuencia, familiaridad y vecindad ortografica. *Psicologica, 14*, 307-317.

Perez, M. A., & Navalon, C. (2003). A set of 290 new pictures standardized for Spanish: Name agreement, image agreement, familiarity, visual complexity and image variability / Normas espaAolas de 290 nuevos dibujos: Acuerdo en la denominacion, concordancia de la imagen, familiaridad, complejidad visual y variabilidad de la imagen. *Psicologica, 24*, 215-241.

Perez, M. A., & Navalon, C. (in press). Objective AoA norms for 175 names in Spanish: Relationships with other psycholinguistic variables, estimated AoA and data from other languages. *European Journal of Cognitive Psychology.*

Pexman, P. M., Lupker, S. J., & Jared, D. (2001). Homophone effects in lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 139-156.

Pierce, J. R., & Karlin, J. E. (1957). Reading rates and the information of a human channel. *Bell Systems Technical Journal, 11*, 497-516.

Pind, J., Jonsdottir, H., Gissurardottir, H., & Jonsson, F. (2000). Icelandic norms for the Snodgrass and Vanderwart (1980) pictures: Name and image agreement, familiarity, and age of acquisition. *Scandinavian Journal Of Psychology, 41*, 41-48.

Pineiro, A., & Manzano, M. (2000). A lexical database for Spanish-speaking children. *Behavior Research Methods Instruments & Computers, 32*, 616-628.

Pineiro, A., Manzano, M., & Reigosa, V. (1999). A standardized set of 257 pictures in Cuban Spanish-speaking children / Estandarizacion de un conjunto de 257 figuras en ninos de habla hispana cubanos. *Cognitiva, 11*, 215-242.

Plaut, D. C. (1997). Structure and function in the lexical system: Insights from distributed models of word reading and lexical decision. *Language And Cognitive Processes, 12*, 765-805.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired reading: Computational principles in quasi-regular domains. *Psychological Review, 103*, 56-115.

Pompeia, S., Miranda, M. C., & Bueno, O. F. A. (2003). Brazilian standardised norms for a set of pictures are comparable with those obtained internationally. *Arquivos de Neuro Psiquiatria, 61*, 916-919.

Rastle, K., & Coltheart, M. (1998). Whammies and double whammies: The effect of length on nonword reading. *Psychonomic Bulletin & Review, 5*, 277-282.

Rastle, K., & Coltheart, M. (1999). Lexical and nonlexical phonological priming in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance, 25*, 461-481.

Rastle, K., & Coltheart, M. (1999). Serial and strategic effects in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance, 25*, 482-503.

Rastle, K., & Davis, M. H. (2002). On the complexities of measuring naming. *Journal of Experimental Psychology: Human Perception and Performance, 28*, 307-314.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*, 372-422.

Rayner, K., & Raney, G. E. (1996). Eye movement control in reading and visual search: Effects of word frequency. *Psychonomic Bulletin & Review, 3*, 245-248.

Rayner, K., & Sereno, S. C. (1994). Eye movements in reading: Psycholinguistic studies. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 57-81). San Diego, CA: Academic Press.

Reynolds, M., & Besner, D. (2004). Neighbourhood Density, Word Frequency, and Spelling-Sound Regularity Effects in Naming: Similarities and Differences Between Skilled Readers and the Dual Route Cascaded Computational Model. *Canadian Journal Of Experimental Psychology-Revue Canadienne De Psychologie Experimentale, 58*, 13-31.

Rice, G. A., & Robinson, D. O. (1975). The role of bigram frequency in the perception of words and nonwords. *Memory & Cognition, 3*, 513-518.

Richardson, J. T. E. (1975). The effect of word imageability in acquired dyslexia. *Neuropsychologia, 13*, 281-288.

Rinsland, J. C. (1945). *A basic vocabulary of elementary school children.* New York: Macmillan.

Roget, P. M. (1911). *Roget's Thesaurus of English Words and Phrases (1911 edition).*

Roodenrys, S., Hulme, C., Alban, J., Ellis, A. W., & Brown, G. D. A. (1994). Effects Of Word-Frequency And Age Of Acquisition On Short-Term- Memory Span. *Memory & Cognition, 22*, 695-701.

Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior, 9*, 487-494.

Rubenstein, H., Lewis, S. S., & Rubenstein, M. A. (1971). Evidence for phonemic recoding in visual word recognition. *Journal of Verbal Learning and Verbal Behavior, 10*, 645-657.

Rubenstein, H., Lewis, S. S., & Rubenstein, M. A. (1971). Homographic entries in the internal lexicon: Effects of systematically and relative frequency of meanings. *Journal of Verbal Learning and Verbal Behavior, 10*, 57-62.

Rubin, D. C. (1980). 51 properties of 125 words: A unit analysis of verbal behavior. *Journal of Verbal Learning and Verbal Behavior, 19*, 736-755.

Rugg, M. D. (1990). Event-related brain potentials disociate repetition effects of high- and low-frequency words. *Memory & Cognition, 18.*

Rumelhart, D. E., & McClelland, J. L. (1982). An Interactive Activation Model Of Context Effects In Letter Perception.2. The Contextual Enhancement Effect And Some Tests And Extensions Of The Model. *Psychological Review, 89*, 60-94.

Sanfeliu, M. C., & Fernandez, A. (1996). A set of 254 Snodgrass-Vanderwart pictures standardized for Spanish: Norms for name agreement, image agreement, familiarity, and visual complexity. *Behavior Research Methods, Instruments and Computers, 28*, 537-555.

Schilling, H. E. H., Rayner, K., & Chumbley, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory & Cognition, 26*, 1270-1281.

Sebastian, N., Marti, M. A., Carreiras, M. G., & Cuetos, F. (2000). *LEXESP, lexico informatizado del Espanol.*Barcelona: Ediciones de la Universitat de Barcelona.

Seidenberg, M. S. (1995). Visual word recognition: An overview. In P. Eimas & J. L. Miller (Eds.), *Handbook of perception and cogntion: Language* (2nd ed.) (Vol. 11, pp. 137-179). New York: Academic Press.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96*, 523-568.

Selfridge, O. G., & Neisser, U. (1960). Pattern recognition by machine. *Scientific American, 203 (Aug)*, 60-68.

Sereno, S. C., Rayner, K., & Posner, M. I. (1998). Establishing a time-line of word recognition: Evidence from eye movements and event-related potentials. *NeuroReport: An International Journal for the Rapid Communication of Research in Neuroscience, 9*, 2195-2200.

Shu, H., Zhang, H., Li, W., Wang, A., & al., e. (1992). A new technique for cognitive experiments: Chinese norm for a set of pictures and a computer program for experiments. *Acta Psychologica Sinica, 24*, 386-392.

Smith, M. A., Cottrell, G. W., & Anderson, K. L. (2001). The early word catches the weights. In T. K. Leen, T. G. Dietterich & V. Tresp (Eds.), *Advances in neural information processing systems 13* (pp. 52-58). Cambridge, MA: MIT Press.

Snodgrass, J. G. (1984). Concepts and their surface representations. *Journal of Verbal Learning and Verbal Behavior, 23*, 3-22.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 174-215.

Snodgrass, J. G., & Yuditsky, T. (1996). Naming times for the Snodgrass and Vanderwart pictures. *Behavior Research Methods Instruments & Computers, 28*, 516-536.

Spieler, D. H., & Balota, D. A. (1997). Bringing computational models of word naming down to the item level. *Psychological Science, 8*, 411-416.

Spreen, O., & Schulz, R. W. (1966). Parameters Of Abstraction Meaningfulness And Pronunciability For 329 Nouns. *Journal Of Verbal Learning And Verbal Behavior, 5*, 459-&.

Stadthagen-Gonzalez, H., Bowers, J. S., & Damian, M. F. (2004). Age-of-acquisition effects in visual word recognition: evidence from expert vocabularies. *Cognition, 93*, B11-B26.

Stadthagen-Gonzalez, H., & Davis, C. J. (in press). The Bristol Norms for Age of Acquisition, Imageability and Familiarity. *Behavior Research Methods, Instruments and Computers.*

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive Science, 29*, 41-78.

Stone, G. O., Vanhoy, M., & Van Orden, G. C. (1997). Perception is a two-way street: Feedforward and feedback phonology in visual word recognition. *Journal of Memory and Language, 36*, 337-359.

Strain, E., Patterson, K., & Seidenberg, M. S. (1995). Semantic Effects In Single-Word Naming. *Journal Of Experimental Psychology-Learning Memory And Cognition, 21*, 1140-1154.

Strain, E., Patterson, K., & Seidenberg, M. S. (2002). Theories of word naming interact with spelling-sound consistency. *Journal Of Experimental Psychology-Learning Memory And Cognition, 28*, 207-214.

Stratton, R. P., Jacobus, K. A., & Brinley, B. (1975). Age-of-acquisition, imagery, familiarity and meaningfulness norms for 543 words. *Behavior Research Methods and Instrumentation, 7*, 1-6.

Taft, M. (1991). *Reading and the Mental Lexicon.* London: Lawrence Erlbaum Associates.

Taylor, R. (1998). Effects of age of acquisition, word frequency, and familiarity on object recognition and naming in dementia. *Perceptual And Motor Skills, 87*, 573-574.

Theios, J. G., & Amrhein, P. C. (1989). Theoretical analysis of the cognitive processing of lexical and pictorial stimuli: reading, naming, and visual and conceptual comparison. *Psychological Review, 96*, 5-24.

Thorndike, E. L. (1921). *The teacher's word book.* New York: Teachers College, Columbia University.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words.* New York: Teahcers College, Columbia University.

Toglia, M. P., & Battig, W. R. (1978). *Handbook of Semantic Word Norms.* New York: Erlbaum.

Treisman, A. M. (1964). Monitoring and storage of irrelevant messages in selective attention. *Journal of Verbal Learning and Verbal Behavior, 3*, 449-459.

Tryk, H. E. (1968). Subjective scaling of word frequency. *American Journal Of Psychology, 81*, 170-177.

Turner, J. E., Valentine, T., & Ellis, A. W. (1998). Contrasting effects of age of acquisition and word frequency on auditory and visual lexical decision. *Memory & Cognition, 26*, 1282-1291.

Ukita, H., Abe, K., & Yamada, J. (1999). Late acquired words in childhood are lost earlier in primary progressive aphasia. *Brain And Language, 70*, 205-219.

Valle-Arroyo, F. (1998). *Normas de imaginabilidad.* Oviedo: Servicio de Publicaciones de la Universidad de Oviedo.

Van Loon-Vervoorn, W. A. (1989). *Eigenshappen van basiswordeu.* Lisse: Swets and Zeitlinger.

Van Petten, C., & Kutas, M. (1992). Interactions between sentence context and word frequency in event-related brain potentials. *Memory & Cognition, 18.*

Van Schagen, I., & al., e. (1983). Names and norms for pictures / Namen en normen voor plaatjes. *Nederlands Tijdschrift voor de Psychologie en haar Grensgebieden, 38*, 236-241.

Vitkovitch, M., & Tyrrell, L. (1995). Sources Of Disagreement In Object Naming. *Quarterly Journal Of Experimental Psychology Section A-Human Experimental Psychology, 48A*, 822-848.

Voyer, D. (2003). Word frequency and laterality effects in lexical decision: Right hemisphere mechanisms. *Brain And Language, 87*, 421-431.

Walker, H. J. (1970). Imagery ratings for 338 nouns. *Behavior Research Methods and Instrumentation, 2*, 165-167.

Walley, A. C., & Metsala, J. L. (1992). Young Childrens Age-Of-Acquisition Estimates For Spoken Words. *Memory & Cognition, 20*, 171-182.

Ward, G., & Maylor, E. A. (2005). Age-related deficits in free recall: The role of rehearsal. *Quarterly Journal Of Experimental Psychology: Human Experimental Psychology, 58*, 98-119.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature, 393*.

Whaley, C. P. (1978). Word-nonword classification time. *Journal of Verbal Learning and Verbal Behavior, 17*, 143-154.

Winnick, W. A., & Daniel, S. A. (1970). Two kinds of response priming in tachistoscopic recognition. *Journal Of Experimental Psychology, 84*, 74-81.

Winters, J. J., Winter, L., & Burger, A. L. (1978). Confidence In Age-Of-Acquisition Estimates And Its Relationship To Childrens Labeling Performance. *Bulletin Of The Psychonomic Society, 12*, 361-364.

Yamazaki, M., Ellis, A. W., Morrison, C. M., & Ralph, M. A. L. (1997). Two age of acquisition effects in the reading of Japanese Kanji. *British Journal Of Psychology, 88*, 407-421.

Zeno, S., Ivens, S., Millard, R., & Duvvuri, R. (1995). *The educator's word frequency guide*:Touchstone Applied Science Associates, Inc.

Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal Of Memory And Language, 47*, 1-29.

Zevin, J. D., & Seidenberg, M. S. (2004). Age-of-acquisition effects in reading aloud: Tests of cumulative frequency and frequency trajectory. *Memory & Cognition, 32*, 31-38.

Ziegler, J. C., Montant, M., & Jacobs, A. M. (1997). The feedback consistency effect in lexical decision and naming. *Journal of Memory and Language, 37*, 533-554.

Ziegler, J. C., Tan, L. H., Perry, C., & Montant, M. (2000). Phonology matters: The phonological frequency effect in written Chinese. *Psychological Science, 11*, 234-238.